564227_1.DOC
MLW/AA/saj
August 3, 2005

PATENT APPLICATION
Attorney's Docket No.: 200304450-1 (0918.2033-000)

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

Applicants:	Simon Kasif, Beth T. Logan, Pedro J. Moreno and Baris E. Suzek		
Application No.:	09/724,269	Group:	1631
Filed:	November 28, 2000	Examiner:	Mary K. Zeman
Confirmation No.:	7893		

For: COMPUTER METHOD AND APPARATUS FOR UNIFORM REPRESENTATION OF GENOME SEQUENCES

CERTIFICATE OF MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service with sufficient postage as First Class Mail in an envelope addressed to Commissioner for Patents, P.O. Box 1450, Alexandria, VA 22313-1450 on 8-8-05 Sandra Jammal
Date Signature

Sandra Jammal
Typed or printed name of person signing certificate

DECLARATION OF BETH T. LOGAN AND PEDRO J. MORENO UNDER 37 C.F.R. 1.131

Commissioner of Patents and Trademarks
P.O. Box 2327
Alexandria, VA 22202

pedro

Sir:

We, Beth T. Logan and Pedro J. Moreno declare and state that:

1. We are the co-inventors of the above-identified application.

2. The invention described and claimed in the above-referenced application was conceived at least as early as May 2000, at least three months prior to August 22, 2000, which is the earliest priority date of the U.S. Pat. No. 6,834,239 to Lobanov *et al.*, cited by the Examiner as a basis of the rejection under 35 U.S.C. §102(e). The invention was reduced to practice by July 5, 2000. The conception, reduction to practice and diligence from conception to reduction to practice are evidenced by the attached Exhibits A through J which are described below in an account of the development of the invention to reduction to practice.

3. In late May 2000, Simon Kasif, Beth T. Logan, and Pedro J. Moreno, then all employees of COMPAQ Computer Corporation, subsequently acquired by Hewlett-Packard, and Baris E. Suzek, an intern at COMPAQ Computer Corporation during the Summer of 2000, convened to discuss a novel approach to protein classification whereby proteins would be represented by combining small sequences.

4. On June 9, 2000, Mr. Suzek recorded on the source controlled internal website, hereinafter, "the Website", that "we will try to find a novel approach to protein classification which will help biologists in finding: functional properties of proteins[,] structural properties of proteins[, and] evolutionary properties of proteins [...] ." Mr. Suzek also recorded that the project plan included "[d]evelop[ing] a tool to find the amino acid sequences (presumably short in length) in the proteins that will help to classify them. Ideally, the tool will try to find the short sequences that best matches with the HMM [Hidden Markov Models] in a given database." A print-out of the Website as of June 9, 2000 is presented as Exhibit A. The relevant portions are highlighted.

5. Between June 9, 2000 and June 12, 2000 Mr. Kasif suggested studying the existing tools for sequence analysis and classification such as HMMER and BLIMPS. HMMER is a freely distributable software for protein sequence analysis using Hidden Markov Models (HMM), available from Washington University in St. Louis, Missouri, at the URL http://hmmer.wustl.edu/. BLIMPS (BLocks IMProved Searcher) is a searching

pedro

tool for BLOCKS database. The most current version of this freeware is available at the URL

http://bioweb.pasteur.fr/seqanal/motif/blimps-uk.html.

BLOCKS is a database of multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins. It is maintained by Pittsburgh Supercomputing Center and is accessible through the URL

http://www.psc.edu/general/software/packages/blocks/blocks.html.

Mr. Kasif pointed out that it may be possible that the relationship of the segments of BLOCKS to the proteins could be analogous to the relationship of the phonemes to words. The implication of this suggestion is that protein sequences may be generated from the segments of BLOCKS. By June 12, 2000, Mr. Suzek recorded on the Website:

> The consensus sequences of BLOCKS will be searched against PFAM to see if there is [*sic*] multiple hits per BLOCK [*sic*], which implies that BLOCKS can be building 'blocks' of domains: As a first step consensus seq[uence]s will be generated from BLOCKS database.

(PFAM, a Protein FAMilies database of alignments and HMMs, is a large collection of multiple sequence alignments and hidden Markov models covering many common protein domains and families. It is maintained by the Sanger Institute and is accessible through the URL http://www.sanger.ac.uk/Software/Pfam/.) A print-out of the Website as of June 12, 2000 is presented as Exhibit B. The relevant portions are highlighted.

6. During a meeting held on June 13, 2000, an idea of modeling each protein as a concatenation of BLOCKS segments was proposed. A concatenation of the BLOCKS segments led to the idea of converting each protein to a feature vector comprising information about the presence of each BLOCKS segment in a given protein sequence. On June 13, 2000, Mr. Suzek recorded on the Website:

> Model proteins by concatenation of short "base units" separated by junk. This is similar to the PFAM domain idea except the base units are shorter than domains - more like the size of BLOCKS. Ideally model each base unit with a HMM.

On the same date, Mr. Suzek recorded on the Website under the heading Project Progress:

> For each protein in the SCOP database, we will find the BLOCKS occurring in them. And generate a feature vector with the scores of BLOCKS found in them.

(The SCOP (Structural Classification of Proteins) database is a comprehensive ordering of all proteins of known structure, according to their evolutionary and structural relationships. SCOP is accessible at the URLs http://scop.berkeley.edu/ or http://scop.mrc-lmb.cam.ac.uk/scop.) A print-out of the Website as of June 13, 2000 is presented as Exhibit C. The relevant portions are highlighted.

7. By June 19, 2000, Mr. Suzek had generated feature vectors for all proteins in SCOP by scoring these proteins against the segments of the BLOCKS database (*i.e.* by counting the number of times each BLOCKS segments is contained in each SCOP protein). Mr. Suzek posted the generated vectors on the Website. A print-out of the Website as of June 20, 2000 is presented as Exhibit D. (See entry 7 under the heading Project Report. The relevant portions are highlighted.)

8. Following the generation of the feature vectors for a significant number of proteins, the question of classifying the proteins was addressed. On or before June 20, 2000, a brainstorming session was held during which various techniques for classifying multidimensional vectors were discussed. On June 20, 2000, Mr. Suzek recorded on the Website under the heading Brain Storming:

> Given a feature vector whose entries are based on posterior probabilties of blocks, we could use SVD [Singular Value Decomposition] [...] to reduce the dimensionality of these huge vector (as many components as blocks!) and find the "important" components. Once this mapping from high dimension to low dimension is done we can also find natural clusters, use Gaussian modeling, classify etc.

Mr. Suzek further recorded on the Website that support vector machines (SVMs) can be used to classify protein families and that our results to the technique accepted in the art such as BLAST. (See Exhibit D, entries 8 and 9 under the heading Project Progress.)

9. On or before June 26, 2000, Mr. Suzek recorded on the Website the results of the comparison of the protein classification obtained using support vector machines to that

obtained using known methods described in Jaakkola *et al.* "A Discriminative Framework for Detecting Remote Protein Homologies", J. Comp. Biol., Vol. 7, Num. 1/2 (2000). A print-out of the Website as of June 26, 2000 is presented as Exhibit E. See entry 9 under the heading Project Progress. The relevant portions are highlighted. (The abbreviation "FPR" stands for False Positive Rate.)

10. At a meeting that took place on or around June 27, 2000, methods of scoring proteins that contained a given segment of the BLOCKS database more than once were discussed. Two approaches were proposed. The first approach was to add the scores and second approach was to take the maximum of the scores. On or before July 7, 2000, Mr. Suzek made a corresponding entry on the Website. A print-out of the Website as of July 7, 2000 is presented as Exhibit F. See entry 7 under the heading Project Progress. The relevant portions are highlighted.

11. By July 5, 2000, Mr. Suzek reported completion of training the support vector machines embodiment of the invention based on protein classification obtained using methods described in Jaakkola *et al.* "A Discriminative Framework for Detecting Remote Protein Homologies", J. Comp. Biol., Vol. 7, Num. 1/2 (2000). See entry 9 under the heading Project Progress of Exhibit F. The relevant portions are highlighted.

12. In early August 2000, Mr. Suzek prepared a presentation based on the results of his summer internship at Compaq Computer Corp. A print out of this presentation in the PowerPoint format is presented as Exhibit G. Mr. Suzek's presentation was copied to the Website on August 11, 2000.

13. By August 3, 2000, we commenced drafting an invention disclosure and on August 21, 2000, the final version of the invention disclosure was sent to Compaq legal counsel. A copy of an email, with attachment, to Mr. R. Reed, an engineering liaison to a legal counsel for Compaq Computer Corp., is presented as Exhibit H. In section 4 of the invention disclosure for (Exhibit H), we report implementation of the invention in software between June 15 and July 31, 2000.

14. On September 15, 2000, Mr. R. Lange, a Legal Counsel for Compaq Computer Corp., contacted Ms. MaryLou Wakimura, a principal at the law firm of Hamilton, Brook, Smith & Reynolds with a request to prepare the patent application based on the research work described above. A copy of the email to Ms. Wakimura is presented as Exhibit I.

15. On October 5, 2000 we met with Ms. Wakimura to discuss drafting the patent application. A copy of the email from Ms. Logan to Ms. Wakimura scheduling this meeting is presented as Exhibit J.

16. Through October and November 2000, Ms. Wakimura produced a patent application which was filed in the USPTO on November 11, 2000 as evidenced by the present subject patent application.

17. I hereby acknowledge that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code and that such willful false statements may jeopardize the validity of the application or any patent issued thereon.

BETH T. LOGAN

Date

Pedro Moreno

PEDRO J. MORENO

August 5, 2005.

Date

pedro

Index of Exhibits

Exhibit A: A print-out of the source-controlled internal Compaq Computer Corp. website as of June 9, 2000.

Exhibit B: A print-out of the source-controlled internal Compaq Computer Corp. website as of June 12, 2000.

Exhibit C: A print-out of the source-controlled internal Compaq Computer Corp. website as of June 13, 2000.

Exhibit D: A print-out of the source-controlled internal Compaq Computer Corp. website as of June 20, 2000.

Exhibit E: A print-out of the source-controlled internal Compaq Computer Corp. website as of June 26, 2000.

Exhibit F: A print-out of the source-controlled internal Compaq Computer Corp. website as of July 7, 2000.

Exhibit G: A print out of the presentation slides by Mr. Suzek.

Exhibit H: A copy of an email dated August 21, 2000, from Ms. Logan to Mr. Reed and the attached invention disclosure.

Exhibit I: A copy of an email dated September 15, 2000, from Ms. Wakimura to Mr. Lange regarding preparation of the subject patent application.

Exhibit J: A copy of an email dated September 22, 2000, from Ms. Logan to Ms. Wakimura regarding scheduling a meeting to discuss the subject patent application.

Using Speech Techniques in Computational Biology

by Baris Suzek supervised by Beth Logan and Simon Kasif

In this project we will try to find a novel approach to protein classification which will help biologists in finding :

- functional properties of proteins
- structural properties of proteins
- evolutionary properties of proteins

and lots of other things that we can currently can't imagine. To achieve this, we are planning to use well established speech recognition techniques such as hidden markov models (HMM's).

1. Project Plan

Develop a tool to find the amino acid sequences (presumably short in length) in the proteins that will help to classify them. Ideally, the tool will try to find the short sequences that best matches with the HMM models in a given database. A major amount of modification will be made to an existing HMM tool, namely "CALISTA", to be used in this project and for future Bioinformatic projects.

2. Ideas

It seems that people don't need more than about 10s-30s of a song to classify it, so the features should capture about that much of the signal. Multiple segments from the same song may not all be close together in the feature space, so presumably outliers (which may be silence, boundaries between differing regions, etc.) should be ignored by the model.

Process

Although the high-level genre distinction can be considered somewhat 'labeled' by the organization of a database of music, we hope to automate the clustering process to some degree in terms of a distance metric. This would hopefully allow us to measure subclasses within the major classes, etc.

Features

Major components: overall rhythmic structure, types of instrumentation, what else?

These should include temporal information, e.g. beat spectrum (Foote) -- actually, some features extracted from the similarity matrix might be more appropriate, expect to try PCA.

We also want spectral info which should be associated with instrumentation, etc. however, it should be normalized in some way to correct for pitch.. perhaps looking at broad distributions of spectral energy? Questions: are MFCC useful? (clearly we want auditory weighting of spectrum, but does cepstral processing decorrelate in useful ways for this task?); do we want some information about excitation statistics (note Dubnov & Tishby paper, correspond to instrument type)?

EXHIBIT
A

Model

Gaussian classifiers have been used, as well as simple distance measures (Muscle Fish paper) or MAP classifiers. We are interested in using a Support Vector Machine classifier.

Assuming that the classifier is used on individual segments from a song, some mechanism will be needed to 'vote' or average among the estimates generated.

3. Project Progress

4. Software Documentation

5. Bibliography

6. Links

Using Speech Techniques in Computational Biology

by Baris Suzek supervised by Beth Logan and Simon Kasif

In this project we will try to find a novel approach to protein classification which will help biologists in finding :

- functional properties of proteins
- structural properties of proteins
- evolutionary properties of proteins

and lots of other things that we can currently can't imagine. To achieve this, we are planning to use well established speech recognition techniques such as hidden markov models (HMM's).

1. Project Plan

Develop a tool to find the amino acid sequences (presumably short in length) in the proteins that will help to classify them. Ideally, the tool will try to find the short sequences that best matches with the HMM models in a given database. A major amount of modification will be made to an existing HMM tool, namely "CALISTA", to be used in this project.

2. Preliminary Work

We started by investigating the existing approaches to the protein classification. So far, we examined two tools/databases: HMMER/PFAM and BLIMPS/BLOCKS.

Hmmer and PFAM

Overview

HMMER is a package that have following programs:

- *hmmalign* : align sequences to an existing model
- *hmmbuild:* build a model from multiple sequence alignment
- *hmmcalibrate:* increase the sensitivity of database searches
- *hmmconvert:* convert model file into different formats
- *hmmemit:* emit sequences probabilitly from a profile HMM
- *hmmfetch:* get an existing model from an HMM database
- *hmmindex:* index an HMM database
- *hmmpfam:* search HMM database for matches to query sequence

EXHIBIT
B

- *hmmsearch:* search a sequence database for matches to a query sequences

The local copy of HMMER is available at "/tmp_mnt/mustang/udir4/baris/hmmer". A detailed manual to use can be found in the directory "/tmp_mnt/mustang/udir4/baris/hmmer/Userguide".

PFAM is a database of domain families. The domains are grouped into two in the database PFAM-A and PFAM-B. The PFAM-A domains are the ones that are well characterized domains with high quality alignments e.g ig, GP120 or GP41. The PFAM-B domains are obtained by clustering and aligning the sequences after removal of PFAM-A domains. The major goal in of PFAM-B is introducing the largest PFAM-B families to PFAM-A in the future. For each domain family there is a profile HMM model generated using "hmmerbuild" program of HMMER package. In the next section we will mention the generation process more detailed.

Database	**Families**	**Sequences**	**Residues**
PFAM-A	2128	181068	42018555
PFAM-B	42357	103709	24762358

Table 1: The size of PFAM databases

Model Generation

To generate a model following steps are followed:

1. Generate a multiple sequence alignment for the domain of interest. The "seed alignments" are selected for these purpose which is a subset of all the proteins containing the domain of interest.CLUSTALW can be used for the multiple sequence alignment. A local copy of this tool is available in the directory "/tmp_mnt/mustang/udir4/baris/clustalw"

2. Having the multiple sequence alignment use "hmmbuild" to generate model as:

   hmmbuild *hmmfilename alignedseqfilename*

Information about the options of "hmmbuild" can be found in the userguide.

Finding Domains in a Protein

There are two programs for this purpose "hmmsearch" and "hmmpfam". The first one is used for searching one domain/model in a given protein and used as:

hmmsearch *hmmfilename proteinfile*

The second tool "hmmpfam" is used for searching a given protein sequence for all the domains in PFAM and used as:

hmmpfam *pfamfilename proteinfile*

Design Issues in HMMER

So far we find out following design issues(under construction):

1. The model architecture it is using is called Plan7:

where:

- I's are the insertion states
- D's are deletion states
- M's are match states
- J , C and N are random sequence states that uses preset emission probabilities learned from data and each transition leaving these states are equally likely. J serves for the possibility of multiple occurrences of the same domain in the protein.

2. Given the aligned sequences the transition and emission probabilities are computed by counting the aligned columns. The most important part of this process is deciding whether a column belongs to an insertion or match state. Once this decision made the deletion transition probabilities are calculated by counting the gaps in match states (there is no transition from insert states to delete states) .

3. As expected, some transition and emission probabilities may not be seen in the training alignment. Hence, they have zero probabilities. One approach to resolve this problem is pseudocounts. However HMMER uses a more complicated approach that is the mixtures of Dirichlet distributions. The idea behind this approach is the application of different sets of pseudocount priors for different types of alignment environments. One set might be relevant for loop environments, one for small residue environments etc.

4. Each sequence in the alignment has a weight based on a tree connecting sequences in which the branch lengths indicate the relative degrees of divergence of each edge. The default algorithm used is GSC(Gerstein, Sonnhammer & Chothia)

5. Before the model is constructed the number of states in the model is estimated as the weighted average length of sequences in the "seed alignment".

Blimps and BLOCKS

Overview

Blimps is a searching tool for BLOCKS database , that scores a sequence against blocks or a block against sequences. BLOCKS is a database of multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins.Currently there are 10783 blocks in this database

A local version of Blimps is available at "/tmp_mnt/mustang/udir4/baris/blimps/bin" and BLOCKS at "/tmp_mnt/mustang/udir4/baris/blocks/"

Using Blimps

Blimps is used as follows:

blimps configuration_file

where configuration file contains all the arguments to run the program. A simple configuration file for querying a sequence against all BLOCKS database looks like:

```
SQ   sample.seq
DB   blocks.dat
OU   sample.out
```

where the SQ is the tag for query sequence, DB is the tag for blocks database and OU is the one for output file. A more detailed description of file can be found here.

3. Project Progress

Following tasks are in progress or accomplished:

1. Investigation of protein classification tool: Still ongoing....

2. For each protein in the database we generated a list of PFAM-A/B domains found in them: Done

3. For each domain ,based on the Swissprot database, we calculated conditional probability of PFAM-A domains X and Y being in the same protein, where X and Y are any domains.

4. The consensus sequences of BLOCKS will be searched against PFAM to see if there is multiple hits per BLOCK, which implies that BLOCKS can be building 'blocks' of domains: As a first step consensus seqs will be generated from BLOCKS database.

5. For each protein in the database we will generate a list of BLOCKS found in them: Still on decision phase

6. SCOP database will be studied to find out if there can be found some short sequences(domains) that may be used to structural protein family classification

4. Bibliography

Using Speech Techniques in Computational Biology

by Baris Suzek supervised by Beth Logan and Simon Kasif

In this project we will try to find a novel approach to protein classification which will help biologists in finding :

- functional properties of proteins
- structural properties of proteins
- evolutionary properties of proteins

To achieve this, we are planning to use well established speech recognition techniques such as hidden markov models (HMM's).

1. Project Plan

Model proteins by concatenation of short "base units" separated by junk. This is similar to the PFAM domain idea except the base units are shorter than domains - more like the size of BLOCKS. Ideally model each base unit with a HMM. The hope is that by using smaller units than domains, we can train better models since more more data will be available for each model. For example, in speech recognition, great gains have been made by modeling phones instead of words. For computational biology, this is an unproven statement, but Simon has some evidence that the size of a base unit should be more like 5-55 aa's (avg. 26 block size) than 9-1326 aa's (avg. 240 domain size).

2. Preliminary Work

We started by investigating the existing approaches to the protein classification. So far, we examined two tools/databases: HMMER/PFAM and BLIMPS/BLOCKS.

Hmmer and PFAM

Overview

HMMER is a package that have following programs:

- *hmmalign* : align sequences to an existing model
- *hmmbuild:* build a model from multiple sequence alignment
- *hmmcalibrate:* increase the sensitivity of database searches
- *hmmconvert:* convert model file into different formats
- *hmmemit:* emit sequences probabilitly from a profile HMM
- *hmmfetch:* get an existing model from an HMM database

- *hmmindex:* index an HMM database
- *hmmpfam:* search HMM database for matches to query sequence
- *hmmsearch:* search a sequence database for matches to a query sequences

The local copy of HMMER is available at "/tmp_mnt/mustang/udir4/baris/hmmer". A detailed manual to use can be found in the directory "/tmp_mnt/mustang/udir4/baris/hmmer/Userguide".

PFAM is a database of domain families. The domains are grouped into two in the database PFAM-A and PFAM-B. The PFAM-A domains are the ones that are well characterized domains with high quality alignments e.g ig, GP120 or GP41. The PFAM-B domains are obtained by clustering and aligning the sequences after removal of PFAM-A domains. The major goal in of PFAM-B is introducing the largest PFAM-B families to PFAM-A in the future. For each domain family there is a profile HMM model generated using "hmmerbuild" program of HMMER package. In the next section we will mention the generation process more detailed.

Database	**Families**	**Sequences**	**Residues**
PFAM-A	2128	181068	42018555
PFAM-B	42357	103709	24762358

Table 1: The size of PFAM databases

Model Generation

To generate a model following steps are followed:

1. Generate a multiple sequence alignment for the domain of interest. The "seed alignments" are selected for these purpose which is a subset of all the proteins containing the domain of interest.CLUSTALW can be used for the multiple sequence alignment. A local copy of this tool is available in the directory "/tmp_mnt/mustang/udir4/baris/clustalw"

2. Having the multiple sequence alignment use "hmmbuild" to generate model as:

   hmmbuild *hmmfilename alignedseqfilename*

Information about the options of "hmmbuild" can be found in the userguide.

Finding Domains in a Protein

There are two programs for this purpose "hmmsearch" and "hmmpfam". The first one is used for searching one domain/model in a given protein and used as:

hmmsearch *hmmfilename proteinfile*

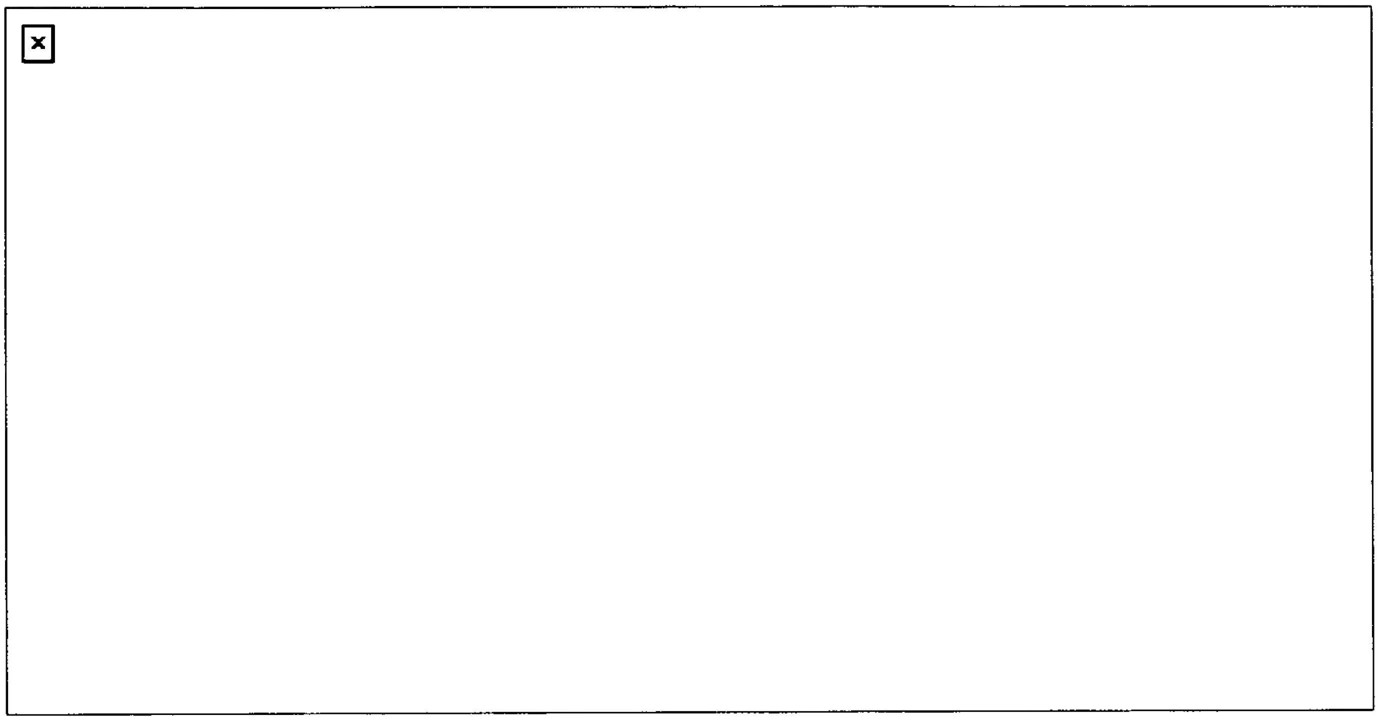
The second tool "hmmpfam" is used for searching a given protein sequence for all the domains in PFAM and used as:

hmmpfam *pfamfilename proteinfile*

Design Issues in HMMER

So far we find out following design issues(under construction):

1. The model architecture it is using is called Plan7:



where:

- I's are the insertion states
- D's are deletion states
- M's are match states
- J , C and N are random sequence states that uses preset emission probabilities learned from data and each transition leaving these states are equally likely. J serves for the possibility of multiple occurrences of the same domain in the protein.

2. Given the aligned sequences the transition and emission probabilities are computed by counting the aligned columns. The most important part of this process is deciding whether a column belongs to an insertion or match state. Once this decision made the deletion transition probabilities are calculated by counting the gaps in match states (there is no transition from insert states to delete states) .

3. As expected, some transition and emission probabilities may not be seen in the training alignment. Hence, they have zero probabilities. One approach to resolve this problem is pseudocounts. However HMMER uses a more complicated approach that is the mixtures of Dirichlet distributions. The idea behind this approach is the application of different sets of pseudocount priors for different types of alignment environments. One set might be relevant for loop environments, one for small residue environments etc.

4. Each sequence in the alignment has a weight based on a tree connecting sequences in which the branch lengths indicate the relative degrees of divergence of each edge. The default algorithm

used is GSC(Gerstein, Sonnhammer & Chothia)

5. Before the model is constructed the number of states in the model is estimated as the weighted average length of sequences in the "seed alignment".

Blimps and BLOCKS

Overview

Blimps is a searching tool for BLOCKS database , that scores a sequence against blocks or a block against sequences. BLOCKS is a database of multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins.Currently there are 10783 blocks in this database

A local version of Blimps is available at "/tmp_mnt/mustang/udir4/baris/blimps/bin" and BLOCKS at "/tmp_mnt/mustang/udir4/baris/blocks/"

Using Blimps

Blimps is used as follows:

```
blimps configuration_file
```

where configuration file contains all the arguments to run the program. A simple configuration file for querying a sequence against all BLOCKS database looks like:

```
SQ   sample.seq
DB   blocks.dat
OU   sample.out
```

where the SQ is the tag for query sequence, DB is the tag for blocks database and OU is the one for output file. A more detailed description of file can be found here.

3. Project Progress

Following tasks are in progress or accomplished:

1. Investigation of protein classification tool: Still ongoing....
2. For each protein in the database we generated a list of PFAM-A/B domains found in them.
3. For each domain ,based on the Swissprot database, we calculated conditional probability of PFAM-A domains X and Y being in the same protein, where X and Y are any domains.
4. The consensus sequences of BLOCKS will be searched against PFAM to see if there is multiple hits per BLOCK, which implies that BLOCKS can be building 'blocks' of domains: As a first step consensus seqs will be generated from BLOCKS database. (Waiting for the results)
5. For each protein in the database we will generate a list of BLOCKS found in them: Still on decision phase

6. SCOP database will be studied to find out if there can be found some short sequences(domains) that may be used to structural protein family classification

7. For each protein in the SCOP database, we will find the BLOCKS occurring in them. And generate a feature vector with the scores of BLOCKS found in them.

4. Brain Storming

5. Bibliography

1. S. Henikoff, J.Henikoff "Protein family classification based on searching a database of blocks", Genomics,1994, 19, 97-107
2. E. Sonnhammer ,S. Eddy, E. Birney "Pfam: multiple sequence alignments and HMM-profiles of protein domains", Nucleic Acids Research,1998,26,320-322
3. E. Sonnhammer ,S. Eddy, R. Durbin "Pfam: a comprehensive database of protein domain families based on seed alignments",Proteins,1997,28,405-420
4. A. Bateman, E. Birney, R. Durbin "The Pfam protein families database", Nucleic Acids Research,2000,28,263-266
5. K. Sjolander, K. Karplus, M. Brown "Dirichlet Mixtures: a method for improving detection of weak but significant protein sequence homology"
6. S. Eddy "Profile hidden Markov models",???
7. R. Durbin, S. Eddy, A. Krogh, G. Mitchison "Biological sequence analysis",Cambridge University Press,1998

Using Speech Techniques in Computational Biology

by Baris Suzek supervised by Beth Logan and Simon Kasif

In this project we will try to find a novel approach to protein classification which will help biologists in finding :

- functional properties of proteins
- structural properties of proteins
- evolutionary properties of proteins

To achieve this, we are planning to use well established speech recognition techniques such as hidden markov models (HMM's).

1. Project Plan

Model proteins by concatenation of short "base units" separated by junk. This is similar to the PFAM domain idea except the base units are shorter than domains - more like the size of BLOCKS. Ideally model each base unit with a HMM. The hope is that by using smaller units than domains, we can train better models since more data will be available for each model. For example, in speech recognition, great gains have been made by modeling phones instead of words. For computational biology, this is an unproven statement, but Simon has some evidence that the size of a base unit should be more like 5-55 aa's (avg. 26 block size) than 9-1326 aa's (avg. 240 domain size).

2. Preliminary Work

We started by investigating the existing approaches to the protein classification. So far, we examined three tools/databases: HMMER/PFAM , BLIMPS/BLOCKS and SCOP.

Hmmer and PFAM

Overview

HMMER is a package that have following programs:

- *hmmalign* : align sequences to an existing model
- *hmmbuild:* build a model from multiple sequence alignment
- *hmmcalibrate:* increase the sensitivity of database searches
- *hmmconvert:* convert model file into different formats
- *hmmemit:* emit sequences probability from a profile HMM
- *hmmfetch:* get an existing model from an HMM database

EXHIBIT
D

- *hmmindex:* index an HMM database
- *hmmpfam:* search HMM database for matches to query sequence
- *hmmsearch:* search a sequence database for matches to a query sequences

The local copy of HMMER is available at "/tmp_mnt/mustang/udir4/baris/hmmer". A detailed manual to use can be found in the directory "/tmp_mnt/mustang/udir4/baris/hmmer/Userguide".

PFAM is a database of domain families. The domains are grouped into two in the database PFAM-A and PFAM-B. The PFAM-A domains are the ones that are well characterized domains with high quality alignments e.g. ig, GP120 or GP41. The PFAM-B domains are obtained by clustering and aligning the sequences after removal of PFAM-A domains. The major goal in of PFAM-B is introducing the largest PFAM-B families to PFAM-A in the future. For each domain family there is a profile HMM model generated using "hmmerbuild" program of HMMER package. In the next section we will mention the generation process more detailed.

Database	**Families**	**Sequences**	**Residues**
PFAM-A	2128	181068	42018555
PFAM-B	42357	103709	24762358

Table 1: The size of PFAM databases

Model Generation

To generate a model following steps are followed:

1. Generate a multiple sequence alignment for the domain of interest. The "seed alignments" are selected for these purpose which is a subset of all the proteins containing the domain of interest.CLUSTALW can be used for the multiple sequence alignment. A local copy of this tool is available in the directory "/tmp_mnt/mustang/udir4/baris/clustalw"

2. Having the multiple sequence alignment use "hmmbuild" to generate model as:

   hmmbuild *hmmfilename alignedseqfilename*

Information about the options of "hmmbuild" can be found in the userguide.

Finding Domains in a Protein

There are two programs for this purpose "hmmsearch" and "hmmpfam". The first one is used for searching one domain/model in a given protein and used as:

hmmsearch *hmmfilename proteinfile*

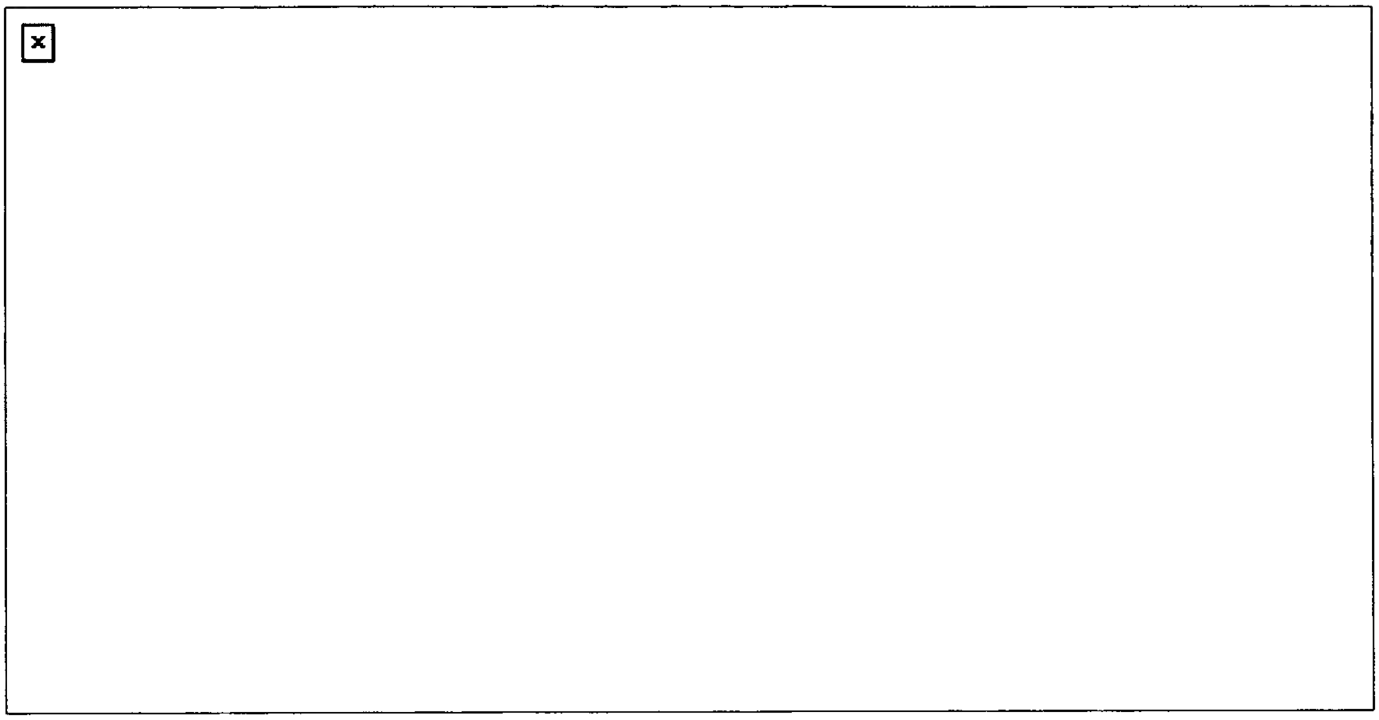
The second tool "hmmpfam" is used for searching a given protein sequence for all the domains in PFAM and used as:

hmmpfam *pfamfilename proteinfile*

Design Issues in HMMER

So far we find out following design issues(under construction):

1. The model architecture it is using is called Plan7:



where:

- I's are the insertion states
- D's are deletion states
- M's are match states
- J , C and N are random sequence states that uses preset emission probabilities learned from data and each transition leaving these states are equally likely. J serves for the possibility of multiple occurrences of the same domain in the protein.

2. Given the aligned sequences the transition and emission probabilities are computed by counting the aligned columns. The most important part of this process is deciding whether a column belongs to an insertion or match state. Once this decision made the deletion transition probabilities are calculated by counting the gaps in match states (there is no transition from insert states to delete states) .

3. As expected, some transition and emission probabilities may not be seen in the training alignment. Hence, they have zero probabilities. One approach to resolve this problem is pseudocounts. However HMMER uses a more complicated approach that is the mixtures of Dirichlet distributions. The idea behind this approach is the application of different sets of pseudocount priors for different types of alignment environments. One set might be relevant for loop environments, one for small residue environments etc.

4. Each sequence in the alignment has a weight based on a tree connecting sequences in which the branch lengths indicate the relative degrees of divergence of each edge. The default algorithm

used is GSC(Gerstein, Sonnhammer & Chothia)

5. Before the model is constructed the number of states in the model is estimated as the weighted average length of sequences in the "seed alignment".

Blimps and BLOCKS

Overview

Blimps is a searching tool for BLOCKS database , that scores a sequence against blocks or a block against sequences. BLOCKS is a database of multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins.Currently there are 10783 blocks in this database

A local version of Blimps is available at "/tmp_mnt/mustang/udir4/baris/blimps/bin" and BLOCKS at "/tmp_mnt/mustang/udir4/baris/blocks/"

Using Blimps

Blimps is used as follows:

blimps configuration_file

where configuration file contains all the arguments to run the program. A simple configuration file for querying a sequence against all BLOCKS database looks like:

```
SQ   sample.seq
DB   blocks.dat
OU   sample.out
```

where the SQ is the tag for query sequence, DB is the tag for blocks database and OU is the one for output file. A more detailed description of file can be found here.

SCOP

SCOP is the structural database of proteins. Basically it contains the domains that plays a role in the structure of proteins. These domains are learned from PDB , which is a database of proteins of which 3-D macromolecular structure data primarily determined experimentally by X-ray crystallography and NMR(Nuclear Magnetic Resonance).

Following table shows some statistics about the current version of SCOP(1.50)

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	127	186	278
All beta proteins	87	154	243
Alpha and beta proteins (a/b)	92	147	300
Alpha and beta proteins (a+b)	159	224	330
Multi-domain proteins	23	23	30
Membrane and cell surface proteins	10	16	18

Small proteins	50	70	97
Total	548	820	1296

For each SCOP member there is an SCOP id like; 1.002.044.001.002.021. The id consists of

- SCOP release number -> 1
- Class number -> 2
- Fold number -> 44
- Super family number ->1
- Family number -> 2
- Protein domain id -> 21

So, in short, there is a hierarchy in SCOP from classes to families. In our experiments we will use SCOP 1.37, which is an older version, because we are planning to compare our results with Jaakola's method [8]. The statistics about the dataset that we will use in our experiments can be found here.

3. Project Progress

Following tasks are in progress or accomplished:

1. Investigation of protein classification tool: Still ongoing....
2. For each protein in the database we generated a list of PFAM-A/B domains found in them.The data is available here.
3. For each domain ,based on the Swissprot database, we calculated conditional probability of PFAM-A domains X and Y being in the same protein, where X and Y are any domains.The data is available here.
4. The consensus sequences of BLOCKS will be searched against PFAM to see if there is multiple hits per BLOCK, which implies that BLOCKS can be building 'blocks' of domains: As a first step consensus seqs will be generated from BLOCKS database. (Waiting for the results)
5. For each protein in the database we will generate a list of BLOCKS found in them: Still on decision phase
6. SCOP database will be studied to find out if there can be found some short sequences(domains) that may be used to structural protein family classification
7. For each protein in the SCOP database, we will find the BLOCKS occurring in them. And generate a feature vector with the scores of BLOCKS found in them. The vectors can be found here.
8. We will run the vector support machine on the feature vectors (obtained by finding blocks in

SCOP domains) .

9. In order to evaluate vector support machines, we will run BLAST searches against SCOP database using the negative test sets used in the experiments done by Jaakola & Haussler.

4. Brain Storming

- Given a feature vector whose entries are based on posterior probabilties of blocks, we could use SVD (aka latent semantic inference LSI) to reduce the dimensionality of these huge vector (as many components as blocks!) and find the "important" components. Once this mapping from high dimension to low dimension is done we can also find natural clusters, use Gaussian modeling, classify etc.

5. Bibliography

1. S. Henikoff, J.Henikoff "Protein family classification based on searching a database of blocks", Genomics,1994, 19, 97-107
2. E. Sonnhammer ,S. Eddy, E. Birney "Pfam: multiple sequence alignments and HMM-profiles of protein domains", Nucleic Acids Research,1998,26,320-322
3. E. Sonnhammer ,S. Eddy, R. Durbin "Pfam: a comprehensive database of protein domain families based on seed alignments",Proteins,1997,28,405-420
4. A. Bateman, E. Birney, R. Durbin "The Pfam protein families database", Nucleic Acids Research,2000,28,263-266
5. K. Sjolander, K. Karplus, M. Brown "Dirichlet Mixtures: a method for improving detection of weak but significant protein sequence homology"
6. S. Eddy "Profile hidden Markov models",???
7. R. Durbin, S. Eddy, A. Krogh, G. Mitchison "Biological sequence analysis",Cambridge University Press,1998
8. T.Jaakkola,M.Diekhans,D.Haussler "A discriminative framework for detecting remote protein homologies"

Using Speech Techniques in Computational Biology

by Baris Suzek supervised by Beth Logan and Simon Kasif

In this project we will try to find a novel approach to protein classification which will help biologists in finding :

- functional properties of proteins
- structural properties of proteins
- evolutionary properties of proteins

To achieve this, we are planning to use well established speech recognition techniques such as hidden markov models (HMM's).

1. Project Plan

Model proteins by concatenation of short "base units" separated by junk. This is similar to the PFAM domain idea except the base units are shorter than domains - more like the size of BLOCKS. Ideally model each base unit with a HMM. The hope is that by using smaller units than domains, we can train better models since more data will be available for each model. For example, in speech recognition, great gains have been made by modeling phones instead of words. For computational biology, this is an unproven statement, but Simon has some evidence that the size of a base unit should be more like 5-55 aa's (avg. 26 block size) than 9-1326 aa's (avg. 240 domain size).

2. Preliminary Work

We started by investigating the existing approaches to the protein classification. So far, we examined three tools/databases: HMMER/PFAM , BLIMPS/BLOCKS and SCOP.

Hmmer and PFAM

Overview

HMMER is a package that have following programs:

- *hmmalign* : align sequences to an existing model
- *hmmbuild:* build a model from multiple sequence alignment
- *hmmcalibrate:* increase the sensitivity of database searches
- *hmmconvert:* convert model file into different formats
- *hmmemit:* emit sequences probability from a profile HMM
- *hmmfetch:* get an existing model from an HMM database

EXHIBIT
E

- *hmmindex:* index an HMM database
- *hmmpfam:* search HMM database for matches to query sequence
- *hmmsearch:* search a sequence database for matches to a query sequences

The local copy of HMMER is available at "/tmp_mnt/mustang/udir4/baris/hmmer". A detailed manual to use can be found in the directory "/tmp_mnt/mustang/udir4/baris/hmmer/Userguide".

PFAM is a database of domain families. The domains are grouped into two in the database PFAM-A and PFAM-B. The PFAM-A domains are the ones that are well characterized domains with high quality alignments e.g. ig, GP120 or GP41. The PFAM-B domains are obtained by clustering and aligning the sequences after removal of PFAM-A domains. The major goal in of PFAM-B is introducing the largest PFAM-B families to PFAM-A in the future. For each domain family there is a profile HMM model generated using "hmmerbuild" program of HMMER package. In the next section we will mention the generation process more detailed.

Database	**Families**	**Sequences**	**Residues**
PFAM-A	2128	181068	42018555
PFAM-B	42357	103709	24762358

Table 1: The size of PFAM databases

Model Generation

To generate a model following steps are followed:

1. Generate a multiple sequence alignment for the domain of interest. The "seed alignments" are selected for these purpose which is a subset of all the proteins containing the domain of interest.CLUSTALW can be used for the multiple sequence alignment. A local copy of this tool is available in the directory "/tmp_mnt/mustang/udir4/baris/clustalw"
2. Having the multiple sequence alignment use "hmmbuild" to generate model as:

   hmmbuild *hmmfilename alignedseqfilename*

Information about the options of "hmmbuild" can be found in the userguide.

Finding Domains in a Protein

There are two programs for this purpose "hmmsearch" and "hmmpfam". The first one is used for searching one domain/model in a given protein and used as:

hmmsearch *hmmfilename proteinfile*

The second tool "hmmpfam" is used for searching a given protein sequence for all the domains in PFAM and used as:

hmmpfam *pfamfilename proteinfile*

Design Issues in HMMER

So far we find out following design issues(under construction):

1. The model architecture it is using is called Plan7:

where:

- I's are the insertion states
- D's are deletion states
- M's are match states
- J , C and N are random sequence states that uses preset emission probabilities learned from data and each transition leaving these states are equally likely. J serves for the possibility of multiple occurrences of the same domain in the protein.

2. Given the aligned sequences the transition and emission probabilities are computed by counting the aligned columns. The most important part of this process is deciding whether a column belongs to an insertion or match state. Once this decision made the deletion transition probabilities are calculated by counting the gaps in match states (there is no transition from insert states to delete states) .

3. As expected, some transition and emission probabilities may not be seen in the training alignment. Hence, they have zero probabilities. One approach to resolve this problem is pseudocounts. However HMMER uses a more complicated approach that is the mixtures of Dirichlet distributions. The idea behind this approach is the application of different sets of pseudocount priors for different types of alignment environments. One set might be relevant for loop environments, one for small residue environments etc.

4. Each sequence in the alignment has a weight based on a tree connecting sequences in which the branch lengths indicate the relative degrees of divergence of each edge. The default algorithm

used is GSC(Gerstein, Sonnhammer & Chothia)

5. Before the model is constructed the number of states in the model is estimated as the weighted average length of sequences in the "seed alignment".

Blimps and BLOCKS

Overview

Blimps is a searching tool for BLOCKS database , that scores a sequence against blocks or a block against sequences. BLOCKS is a database of multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins.Currently there are 10783 blocks in this database

A local version of Blimps is available at "/tmp_mnt/mustang/udir4/baris/blimps/bin" and BLOCKS at "/tmp_mnt/mustang/udir4/baris/blocks/"

Using Blimps

Blimps is used as follows:

blimps configuration_file

where configuration file contains all the arguments to run the program. A simple configuration file for querying a sequence against all BLOCKS database looks like:

```
SQ   sample.seq
DB   blocks.dat
OU   sample.out
```

where the SQ is the tag for query sequence, DB is the tag for blocks database and OU is the one for output file. A more detailed description of file can be found here.

SCOP

SCOP is the structural database of proteins. Basically it contains the domains that plays a role in the structure of proteins. These domains are learned from PDB , which is a database of proteins of which 3-D macromolecular structure data primarily determined experimentally by X-ray crystallography and NMR(Nuclear Magnetic Resonance).

Following table shows some statistics about the current version of SCOP(1.50)

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	127	186	278
All beta proteins	87	154	243
Alpha and beta proteins (a/b)	92	147	300
Alpha and beta proteins (a+b)	159	224	330
Multi-domain proteins	23	23	30
Membrane and cell surface proteins	10	16	18

Small proteins	50	70	97
Total	548	820	1296

For each SCOP member there is an SCOP id like; 1.002.044.001.002.021. The id consists of

- SCOP release number -> 1
- Class number -> 2
- Fold number -> 44
- Super family number ->1
- Family number -> 2
- Protein domain id -> 21

So, in short, there is a hierarchy in SCOP from classes to families. In our experiments we will use SCOP 1.37, which is an older version, because we are planning to compare our results with Jaakola's method [8]. The statistics about the dataset that we will use in our experiments can be found here.

3. Project Progress

Following tasks are in progress or accomplished:

1. Investigation of protein classification tool: Still ongoing....
2. For each protein in the database we generated a list of PFAM-A/B domains found in them.The data is available here.
3. For each domain ,based on the Swissprot database, we calculated conditional probability of PFAM-A domains X and Y being in the same protein, where X and Y are any domains.The data is available here.
4. The consensus sequences of BLOCKS will be searched against PFAM to see if there is multiple hits per BLOCK, which implies that BLOCKS can be building 'blocks' of domains: As a first step consensus seqs will be generated from BLOCKS database. (Waiting for the results)
5. For each protein in the database we will generate a list of BLOCKS found in them: Still on decision phase
6. SCOP database will be studied to find out if there can be found some short sequences(domains) that may be used to structural protein family classification
7. For each protein in the SCOP database, we will find the BLOCKS occurring in them. And generate a feature vector with the scores of BLOCKS found in them. The vectors can be found here.
8. Using the feature vectors we generated for training families, provided by Jaakkola, .we trained

our vector support machine and tested the generated models on the test families provided by Jaakola.

9. We used the scores generated by vector support machine, to compute false positive rates so that we can juxtapose our results with Jaakola's. There were two false positive rates calculated in Jaakola's paper; one for 100% coverage and one for 50% coverage. To calculate 100% coverage FPR, we first set the score threshold so that all the positives in the testing set have scores above it and then we computed the FPR using the false's above this threshold. The calculation of 50% coverage FPR was similar, except this time the threshold was selected so that half of the positives have scores above the threshold. The tables showing 50/100% coverage results for 33 families fan be found here.

10. In order to evaluate vector support machines, we will run BLAST searches against SCOP database using the negative test sets used in the experiments done by Jaakola & Haussler.

11. We will include "homologs" to training set for one test family to see if there is a significant improvement in performance of SVM.

4. Brain Storming

- Given a feature vector whose entries are based on posterior probabilties of blocks, we could use SVD (aka latent semantic inference LSI) to reduce the dimensionality of these huge vector (as many components as blocks!) and find the "important" components. Once this mapping from high dimension to low dimension is done we can also find natural clusters, use Gaussian modeling, classify etc.

5. Bibliography

1. S. Henikoff, J.Henikoff "Protein family classification based on searching a database of blocks", Genomics,1994, 19, 97-107
2. E. Sonnhammer ,S. Eddy, E. Birney "Pfam: multiple sequence alignments and HMM-profiles of protein domains", Nucleic Acids Research,1998,26,320-322
3. E. Sonnhammer ,S. Eddy, R. Durbin "Pfam: a comprehensive database of protein domain families based on seed alignments",Proteins,1997,28,405-420
4. A. Bateman, E. Birney, R. Durbin "The Pfam protein families database", Nucleic Acids Research,2000,28,263-266
5. K. Sjolander, K. Karplus, M. Brown "Dirichlet Mixtures: a method for improving detection of weak but significant protein sequence homology"
6. S. Eddy "Profile hidden Markov models",???
7. R. Durbin, S. Eddy, A. Krogh, G. Mitchison "Biological sequence analysis",Cambridge University Press,1998
8. T.Jaakkola,M.Diekhans,D.Haussler "A discriminative framework for detecting remote protein homologies"

Using Speech Techniques in Computational Biology

by Baris Suzek supervised by Beth Logan and Simon Kasif

In this project we will try to find a novel approach to protein classification which will help biologists in finding :

- functional properties of proteins
- structural properties of proteins
- evolutionary properties of proteins

To achieve this, we are planning to use well established speech recognition techniques such as hidden markov models (HMM's).

1. Project Plan

Model proteins by concatenation of short "base units" separated by junk. This is similar to the PFAM domain idea except the base units are shorter than domains - more like the size of BLOCKS. Ideally model each base unit with a HMM. The hope is that by using smaller units than domains, we can train better models since more data will be available for each model. For example, in speech recognition, great gains have been made by modeling phones instead of words. For computational biology, this is an unproven statement, but Simon has some evidence that the size of a base unit should be more like 5-55 aa's (avg. 26 block size) than 9-1326 aa's (avg. 240 domain size).

2. Preliminary Work

We started by investigating the existing approaches to the protein classification. So far, we examined three tools/databases: HMMER/PFAM , BLIMPS/BLOCKS and SCOP.

Hmmer and PFAM

Overview

HMMER is a package that have following programs:

- *hmmalign* : align sequences to an existing model
- *hmmbuild:* build a model from multiple sequence alignment
- *hmmcalibrate:* increase the sensitivity of database searches
- *hmmconvert:* convert model file into different formats
- *hmmemit:* emit sequences probability from a profile HMM
- *hmmfetch:* get an existing model from an HMM database

- *hmmindex:* index an HMM database
- *hmmpfam:* search HMM database for matches to query sequence
- *hmmsearch:* search a sequence database for matches to a query sequences

The local copy of HMMER is available at "/tmp_mnt/mustang/udir4/baris/hmmer". A detailed manual to use can be found in the directory "/tmp_mnt/mustang/udir4/baris/hmmer/Userguide".

PFAM is a database of domain families. The domains are grouped into two in the database PFAM-A and PFAM-B. The PFAM-A domains are the ones that are well characterized domains with high quality alignments e.g. ig, GP120 or GP41. The PFAM-B domains are obtained by clustering and aligning the sequences after removal of PFAM-A domains. The major goal in of PFAM-B is introducing the largest PFAM-B families to PFAM-A in the future. For each domain family there is a profile HMM model generated using "hmmerbuild" program of HMMER package. In the next section we will mention the generation process more detailed.

Database	**Families**	**Sequences**	**Residues**
PFAM-A	2128	181068	42018555
PFAM-B	42357	103709	24762358

Table 1: The size of PFAM databases

Model Generation

To generate a model following steps are followed:

1. Generate a multiple sequence alignment for the domain of interest. The "seed alignments" are selected for these purpose which is a subset of all the proteins containing the domain of interest.CLUSTALW can be used for the multiple sequence alignment. A local copy of this tool is available in the directory "/tmp_mnt/mustang/udir4/baris/clustalw"

2. Having the multiple sequence alignment use "hmmbuild" to generate model as:

   hmmbuild *hmmfilename alignedseqfilename*

Information about the options of "hmmbuild" can be found in the userguide.

Finding Domains in a Protein

There are two programs for this purpose "hmmsearch" and "hmmpfam". The first one is used for searching one domain/model in a given protein and used as:

hmmsearch *hmmfilename proteinfile*

The second tool "hmmpfam" is used for searching a given protein sequence for all the domains in PFAM and used as:

hmmpfam *pfamfilename proteinfile*

Design Issues in HMMER

So far we find out following design issues(under construction):

1. The model architecture it is using is called Plan7:

where:

- I's are the insertion states
- D's are deletion states
- M's are match states
- J , C and N are random sequence states that uses preset emission probabilities learned from data and each transition leaving these states are equally likely. J serves for the possibility of multiple occurrences of the same domain in the protein.

2. Given the aligned sequences the transition and emission probabilities are computed by counting the aligned columns. The most important part of this process is deciding whether a column belongs to an insertion or match state. Once this decision made the deletion transition probabilities are calculated by counting the gaps in match states (there is no transition from insert states to delete states) .

3. As expected, some transition and emission probabilities may not be seen in the training alignment. Hence, they have zero probabilities. One approach to resolve this problem is pseudocount. However HMMER uses a more complicated approach that is the mixtures of Dirichlet distributions. The idea behind this approach is the application of different sets of pseudocount priors for different types of alignment environments. One set might be relevant for loop environments, one for small residue environments etc.

4. Each sequence in the alignment has a weight based on a tree connecting sequences in which the branch lengths indicate the relative degrees of divergence of each edge. The default algorithm

used is GSC(Gerstein, Sonnhammer & Chothia)

5. Before the model is constructed the number of states in the model is estimated as the weighted average length of sequences in the "seed alignment".

Blimps and BLOCKS

Overview

Blimps is a searching tool for BLOCKS database , that scores a sequence against blocks or a block against sequences. BLOCKS is a database of multiply aligned ungapped segments corresponding to the most highly conserved regions of proteins.Currently there are 10783 blocks in this database

A local version of Blimps is available at "/tmp_mnt/mustang/udir4/baris/blimps/bin" and BLOCKS at "/tmp_mnt/mustang/udir4/baris/blocks/"

Using Blimps

Blimps is used as follows:

blimps configuration_file

where configuration file contains all the arguments to run the program. A simple configuration file for querying a sequence against all BLOCKS database looks like:

```
SQ   sample.seq
DB   blocks.dat
OU   sample.out
```

where the SQ is the tag for query sequence, DB is the tag for blocks database and OU is the one for output file. A more detailed description of file can be found here.

SCOP

SCOP is the structural database of proteins. Basically it contains the domains that plays a role in the structure of proteins. These domains are learned from PDB , which is a database of proteins of which 3-D macromolecular structure data primarily determined experimentally by X-ray crystallography and NMR(Nuclear Magnetic Resonance).

Following table shows some statistics about the current version of SCOP(1.50)

Class	Number of folds	Number of superfamilies	Number of families
All alpha proteins	127	186	278
All beta proteins	87	154	243
Alpha and beta proteins (a/b)	92	147	300
Alpha and beta proteins (a+b)	159	224	330
Multi-domain proteins	23	23	30
Membrane and cell surface proteins	10	16	18

Small proteins	50	70	97
Total	548	820	1296

For each SCOP member there is an SCOP id like; 1.002.044.001.002.021. The id consists of

- SCOP release number -> 1
- Class number -> 2
- Fold number -> 44
- Super family number ->1
- Family number -> 2
- Protein domain id -> 21

So, in short, there is a hierarchy in SCOP from classes to families. In our experiments we will use SCOP 1.37, which is an older version, because we are planning to compare our results with Jaakola's method [8]. The statistics about the dataset that we will use in our experiments can be found here.

3. Project Progress

Following tasks are in progress or accomplished:

1. Investigation of protein classification tool: Still ongoing....
2. For each protein in the database we generated a list of PFAM-A/B domains found in them.The data is available here.
3. For each domain ,based on the Swissprot database, we calculated conditional probability of PFAM-A domains X and Y being in the same protein, where X and Y are any domains.The data is available here.
4. The consensus sequences of BLOCKS will be searched against PFAM to see if there is multiple hits per BLOCK, which implies that BLOCKS can be building 'blocks' of domains: As a first step consensus seqs will be generated from BLOCKS database.
5. For each protein in the database we will generate a list of BLOCKS found in them: Still on decision phase
6. SCOP database will be studied to find out if there can be found some short sequences(domains) that may be used to structural protein family classification
7. For each protein in the SCOP database, we will find the BLOCKS occurring in them. And generate a feature vector with the scores of BLOCKS found in them. There are two approaches used to deal with multiple occurrences of a block in same domain; one is summation of scores of multiple domains and the other one is taking the score of maximum scoring occurrence. The vectors can be found here.

8. Some properties of the feature vectors:
   - Each blocks occurs at least one time in the SCOP domains (No all-zero columns)
   - Each SCOP domain has at least one block entry (No all-zero rows)
   - There are on average 309 nonzero entries for each SCOP domain.
9. Using the feature vector which uses summation approach, we generated for training families, provided by Jaakkola, .we trained our vector support machine and tested the generated models on the test families provided by Jaakola.
10. We used the scores generated by vector support machine, to compute false positive rates so that we can juxtapose our results with Jaakola's. There were two false positive rates calculated in Jaakola's paper; one for 100% coverage and one for 50% coverage. To calculate 100% coverage FPR, we first set the score threshold so that all the positives in the testing set have scores above it and then we computed the FPR using the false's above this threshold. The calculation of 50% coverage FPR was similar, except this time the threshold was selected so that half of the positives have scores above the threshold. The tables showing 50/100% coverage results for 33 families can be found here.
11. In order to evaluate vector support machines, we will run BLAST searches against SCOP database using the negative test sets used in the experiments done by Jaakola & Haussler.
12. We will include "homologs" to training set for one test family to see if there is a significant improvement in performance of SVM. We selected the "Long Chain Cytokines (1.25.1.1)". The rate of false positive for this family was 0.123 in Jaakola's paper. For the approach that sums the scores of multiple block hits to generate feature vectors we achieved a RFP of 0.424 and for the approach that takes max score of of multiple block hits we achieved a RFP of 0.16 . For this particular test set they used 2 positive training sets; families 1.25.1.2 ,1.25.1.3,homologs of family 1.25.1.2 and 1.25.1.2 ,1.25.1.3,homologs of family 1.25.1.3. In our experiments we also used the positive training set consisting of 1.25.1.2 ,1.25.1.3,homologs of both families 1.25.1.3 and 1.25.1.2. Here are the results:

	1.25.1.2+1.25.1.3+Homologs of 1.25.1.2	1.25.1.2+1.25.1.3+Homologs of 1.25.1.3	1.25.1.2+1.25.1.3+Hom of 1.25.1.3 and 1.25.1.2
With feature vector that sums scores for of multiple block hits	0.53	0.58	0.62
With feature vectors that takes max of scores of multiple	0.0865	0.1912	0.0857

block hits			

13. We will redo the experiment mentioned in the items 8 and 9 using the feature vectors that uses the maximization approach.

14. We will plot the FN/FP graphs for each familiy for the experiment that uses summation approach in generation of feature vectors.The graphs can be found here.

15. We will plot the FN/FP graphs for each familiy for the experiment that uses max. approach in generation of feature vectors.The graphs can be found here.

16. Here is a comparison between max and sum approach:

    - For 9 families sum approach and for 15 families max approach did better than (or same with) Jaakola's method

    - For 9 families max approach did worse than sum approach

    - For 2 families where sum approach was doing better than Jaakola's method, max approach did worse

17. By taking a lower threshold in BLIMPS we will regenerate feature vectors for SCOP domains and do the experiments mentioned in items 8-9 again with these feature vectors. The approach to deal with multiple blocks in SCOP domains, we will compare the results we obtained from previous experiments.

4. Brain Storming

- Given a feature vector whose entries are based on posterior probabilities of blocks, we could use SVD (aka latent semantic inference LSI) to reduce the dimensionality of these huge vector (as many components as blocks!) and find the "important" components. Once this mapping from high dimension to low dimension is done we can also find natural clusters, use Gaussian modeling, classify etc.

5. Bibliography

1. S. Henikoff, J.Henikoff "Protein family classification based on searching a database of blocks", Genomics,1994, 19, 97-107
2. E. Sonnhammer ,S. Eddy, E. Birney "Pfam: multiple sequence alignments and HMM-profiles of protein domains", Nucleic Acids Research,1998,26,320-322
3. E. Sonnhammer ,S. Eddy, R. Durbin "Pfam: a comprehensive database of protein domain families based on seed alignments",Proteins,1997,28,405-420
4. A. Bateman, E. Birney, R. Durbin "The Pfam protein families database", Nucleic Acids Research,2000,28,263-266
5. K. Sjolander, K. Karplus, M. Brown "Dirichlet Mixtures: a method for improving detection of weak but significant protein sequence homology"
6. S. Eddy "Profile hidden Markov models",???
7. R. Durbin, S. Eddy, A. Krogh, G. Mitchison "Biological sequence analysis",Cambridge University Press,1998
8. T.Jaakkola,M.Diekhans,D.Haussler "A discriminative framework for detecting remote protein

homologies"

Remote Homology Detection in Proteins using Support Vector Machines with Blocks

Baris Suzek   Beth Logan
Simon Kasif   Pedro Moreno
Compaq Cambridge Research Lab

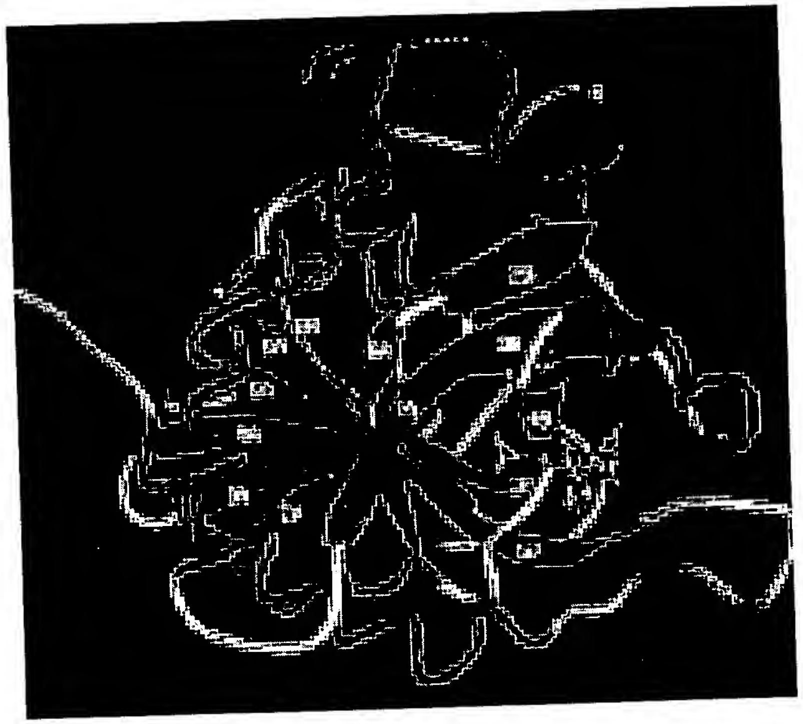
August 2000

Proteins

- Play important functional and structural roles in living organisms
- Annotating proteins with functions and structures ⇒ find homologies
- Lots of putative proteins will be discovered in near future (Human Genome Project)
- Important commercial implications for drug design and discovery

Proteins(2)

>CP0817 MLCATVSGPSFCEAKQQILKSLHLVDIIELRLDLINELDDQELHTLITTAQNPILTFRQH
KEMSTALWIQKLYSLAKLEPKWMDIDVSLPKTALQTIRKSHPKIKLILSYHTDKNEDLDA
IYNEMLATPAEIYKIVLSPENSSEALNYIKKARLLPKPSTVLCMGTHGLPSRVLSPLISN
AMNYAAGISAPQVAPGQPKLEELLSYNYSKLSEKSHIYGLIGDPVDRSISHLSHNFLLSK
LSLNATYIKFPVTIGEVVTFFSAIRDLPFSGLSVTMPLKTAIFDHVDALDASAQLCESIN
TLVFRNQKILGYNTDGEGVAKLLKQKNISVNNKHIAIVGAGGAAKAIAATLAMQGANLHI
FNRTLSSAAALATCCKGKAYPLGSLENFKTIDIIINCLPPEVTFPWRFPPIVMDINTKPH
PSPYLERAQKHGSLIIHGYEMFIEQALLQFALWFPDFLTPESCDSFRNYVKNFMAKV



Domain

- Subsequences of proteins
- Obtained using multiple alignment of proteins of same functional or structural role (human knowledge)

Homologous Proteins

- Share common evolutionary ancestor
- Suggests conserved structure and function
- Implied by similarity between domains or whole sequences
- Can be found by
  - biologists by experiment (tedious and costly)
  - pairwise alignment
  - searching for domains using statistical techniques

Remote homology

- Even for proteins having similar functions(structures) pairwise alignment method fails
- Some success using statistical models built for domains
- We look at a new method to detect remote homologies in proteins

Previous Work

- Jaakola, Diekhans, Haussler
  - HMMs are built for protein domain classes
  - Kernel function(fisher) is derived from posterior probabilities of HMMs
  - SVMs are used to classify
- Strengths
  - Better than using just HMMs or pairwise aligments

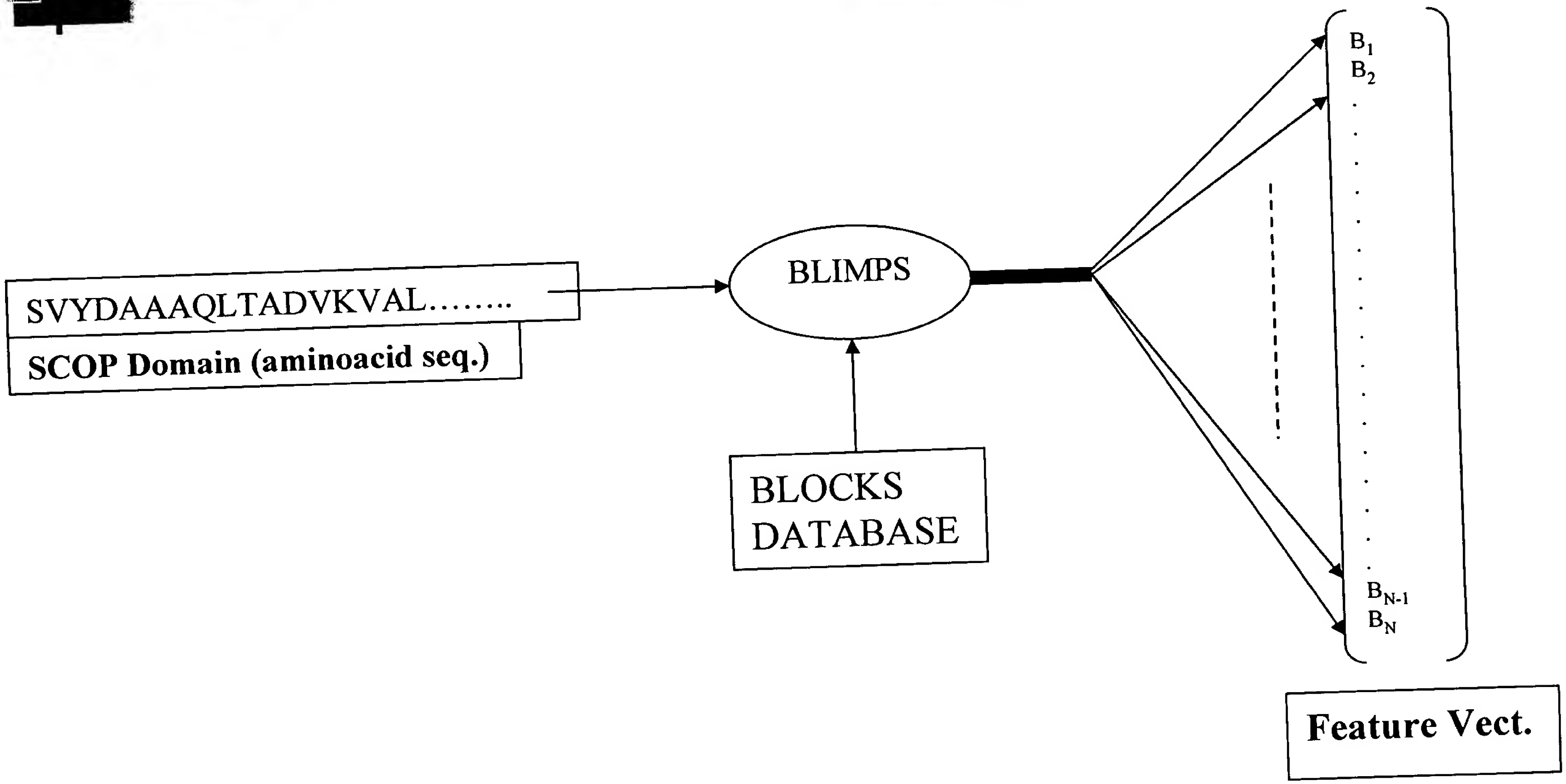
Support Vector Machines

- Binary classification - learns boundary between classes
- Two parameters
  - Capacity - determines generality of model
  - Kernel – Space transformation to simplify classification

Our Technique

- Block (pattern, motif)
  - Short conserved regions among proteins
  - Obtained from multiple alignments of proteins of known function ,structure
  - A database of blocks (BLOCKS) is publicly available
- BLIMPS
  - Generate feature vectors by scoring occurrences of each block in BLOCKS database

Feature Vector Generation



SVYDAAAQLTADVKVAL........

SCOP Domain (aminoacid seq.)

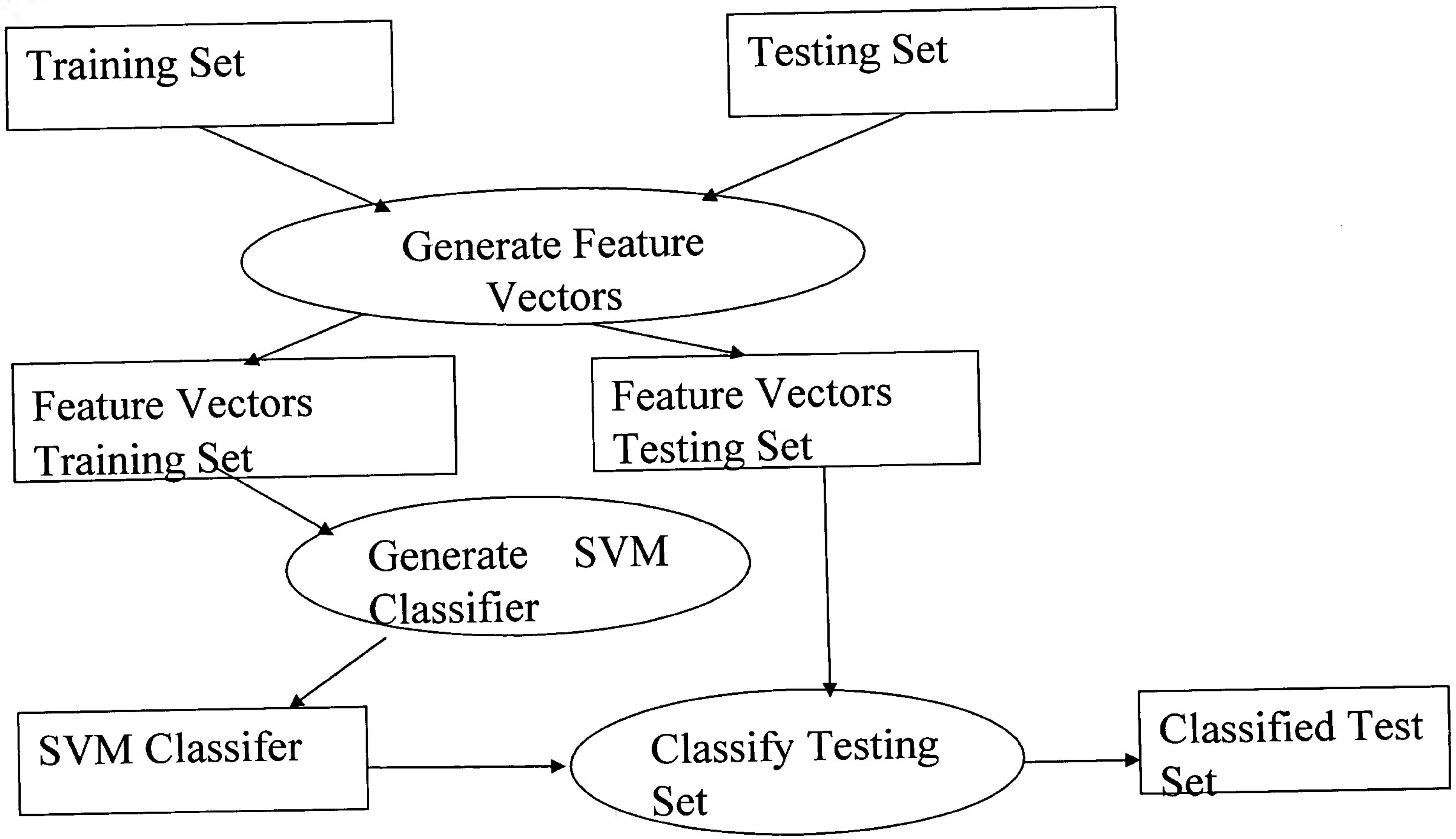
BLIMPS

BLOCKS
DATABASE

B_1
B_2
B_{N-1}
B_N

Feature Vect.

Procedure



Training Set

Testing Set

Generate Feature Vectors

Feature Vectors Training Set

Feature Vectors Testing Set

Generate SVM Classifier

SVM Classifer

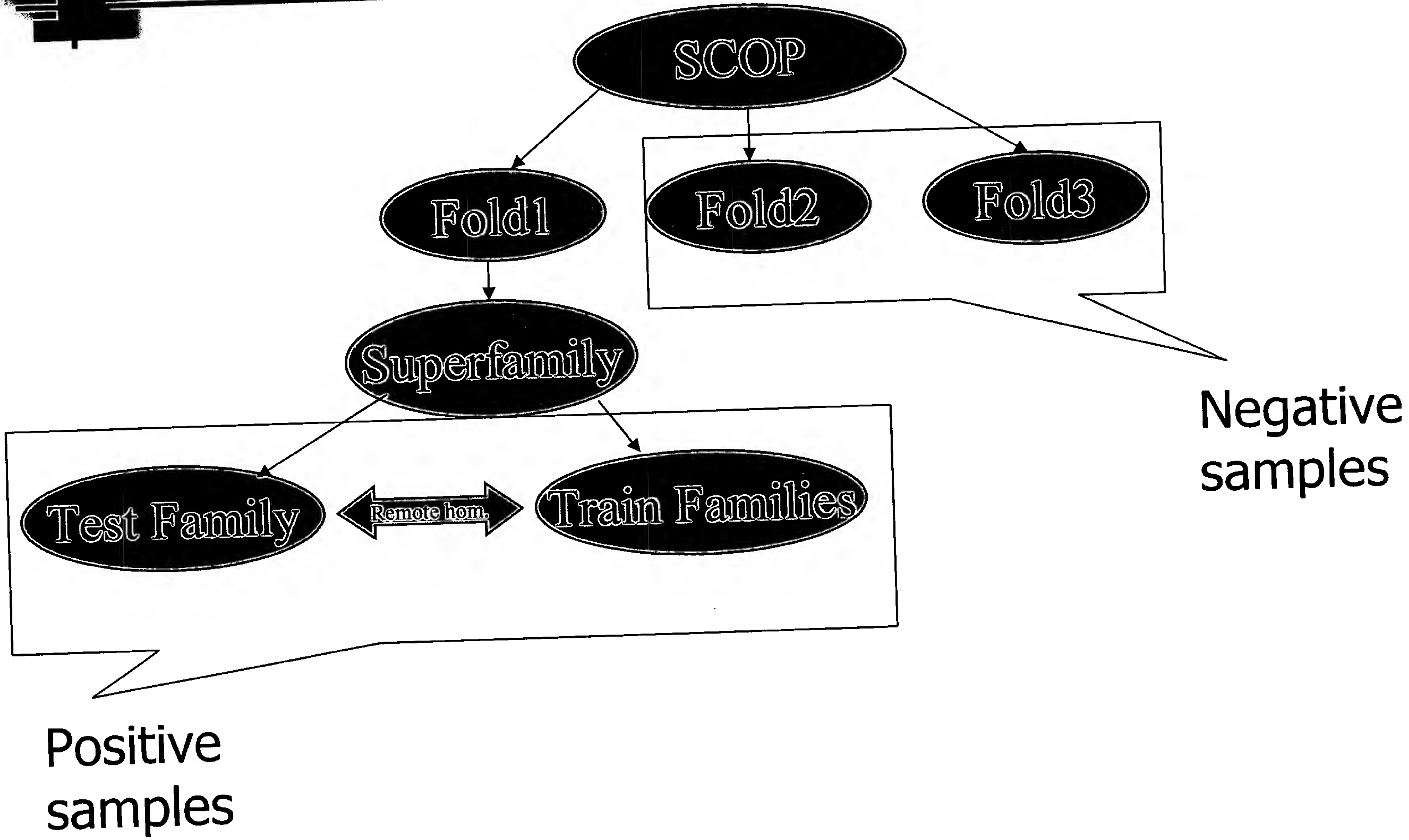
Classify Testing Set

Classified Test Set

Training and Testing Sets

- SCOP(Structural Classification Of Proteins Database)
  - Domains for structure
  - 4-level hierarchy: Class, Fold, Superfamily, Family
- 33 Training and testing sets

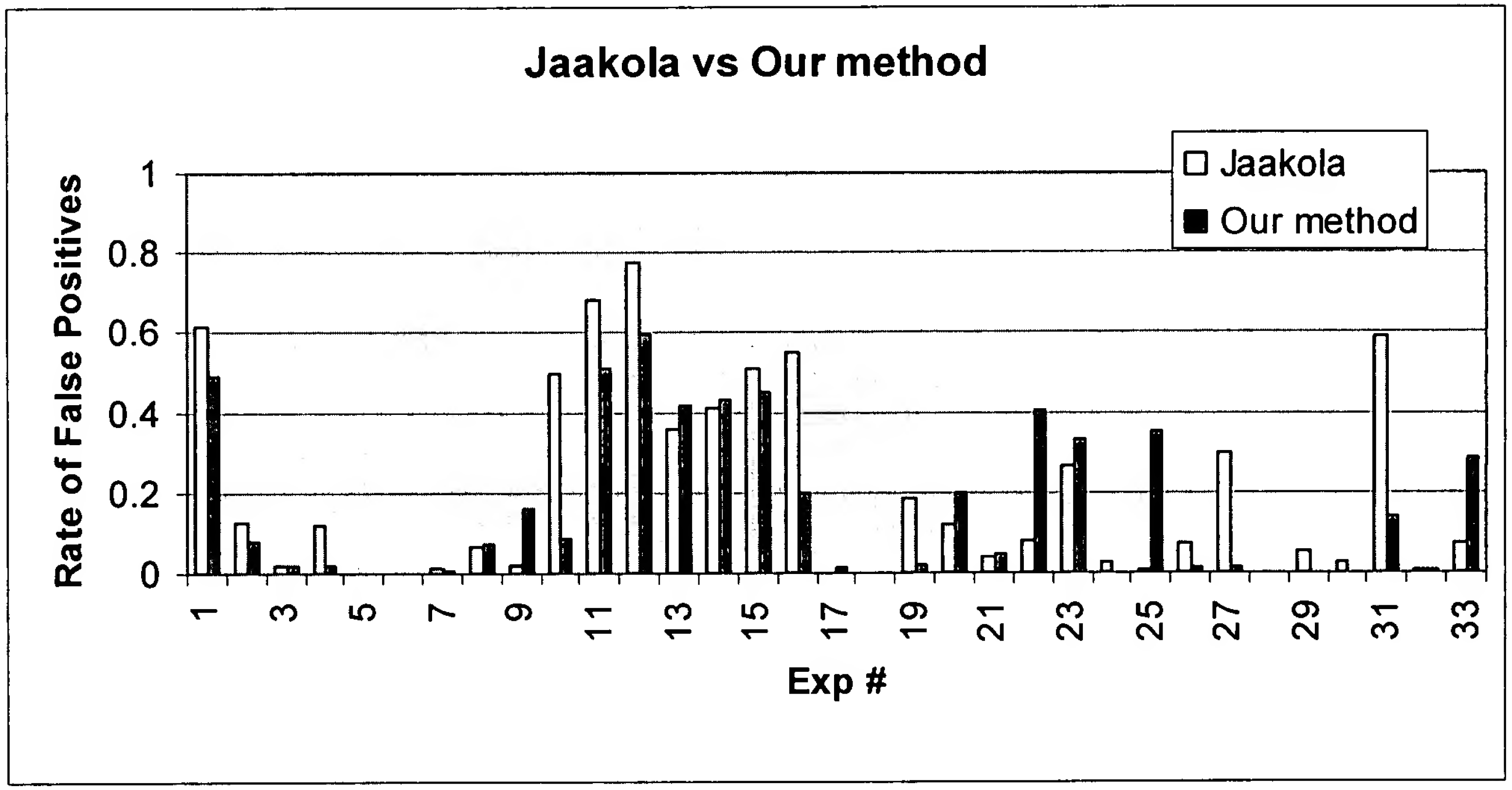
Training and Testing Sets (cont.)



Perfomance

- Rate of false negatives are calculated for 100% coverage (find false positives when all the positive testing are classified correctly by the model)
- 21/33 testing families our model performed better than or the same as Jaakola's method

Performance(continued)



Conclusion and Future Work

- Less complex, better performance
- Blocks are useful for representation of proteins as vectors and can be used for other problems (indexing , clustering)
- Future work
  - In some cases no common blocks among positive training set members since we used blocks from a general database, (BLOCKS) ⇒ build your own structural blocks

Sandra Jammal

From: Beth Logan [IMCEAEX-_O=DIGITAL_OU=CRL_CN=RECIPIENTS_CN=BTL@Compaq.com]
Sent: Monday, August 21, 2000 2:14 PM
To: Reed, Bob
Cc: Beth Logan; Pedro Moreno; Simon Kasif
Subject: Patent disclosure - Technique for Protein and Gene Classification/Clustering/Indexing via a Fixed Dimensional Vector formed using Alignments of Small Motif or Blocks

Bob,

Please find enclose the patent disclosure for "Technique for Protein and Gene Classification/Clustering/Indexing via a Fixed Dimensional Vector formed using Alignments of Small Motif or Blocks" (Kasif/Logan/Moreno/Suzek). As the first written description of the subject matter, I attach the web page from the relevant student project. Since I believe this is one of the first if not the first computational biology patent to be sent to the committee, please do not hesitate to contact us if the subject matter is so unfamiliar that it is difficult to understand.

Also, while I've got your attention I'd like to point out that the new disclosure form is very tedious to use - mostly due to the way the document is formatted. It is almost impossible to navigate the document using the mouse and pictures are difficult to insert. There is also a major problem that the form must be unlocked to check spelling but then check boxes can no longer be checked. (And relocking the form to check check boxes destroys the content of the document!!!) It would be nice if these problems were fixed in the new "research centric" disclosure form.

Thanks
Beth

--
Beth Logan Email: Beth.Logan@compaq.com
Compaq Computer Corporation Ph: +1 617 551 7657
One Cambridge Center Fax: +1 617 551 7650
Cambridge MA 02142 USA WWW: http://www.crl.research.digital.com

EXHIBIT
H

(for Legal Dept. use only)
ID No: _____
Date Received: _____

COMPAQ

COPY

INVENTION DISCLOSURE

PREPARED AND SUBMITTED AT THE REQUEST AND DIRECTION OF AN ATTORNEY
RETURN COMPLETED FORM TO DIANE STRONG VIA E-MAIL TO DIANE.STRONG@COMPAQ.COM

1. **DESCRIPTIVE TITLE OF INVENTION:** Technique for Protein and Gene Classification/Clustering/Indexing via a Fixed Dimensional Vector formed using Alignments of Small Motifs or Blocks

2. **INVENTOR (S):** More than two? ☒ Yes ☐ No (If more than two, use last page.)

A.

Last Name Kasif	Given First Name Simon	Nickname (if any)	Middle Initial/Name
Home Street Address 14 Broadlawn Park		Home Phone	Pager Number
City Chestnut Hill	State MA	Zip 02467	Citizenship USA
Work Phone 617 551 7626	Work Fax 617 551 7650	Mail Code	Employee # 336076
Name of Supervisor R. S. Nikhil		Name of Employer (if NOT an employee of Compaq)	

GROUP	DIVISION
☐ CON	☐ Presario Mobile ☐ Peripherals ☐ Desktop ☐ Internet Services ☐ Con. Adv. Tech.
☐ ESSG	☐ Industry Standard Servers ☐ Storage Products ☐ Telecommunications ☐ High Perf. Server Bus. Unit ☐ Unix Bus. Unit ☐ NonStop™ e-Business ☐ Compaq Services ☐ Professional Services ☐ Alpha ☐ Tandem Bus. Unit
☐ CPCG	☐ Workstations ☐ Desktop ☐ Displays & Peripherals ☐ Portables ☐ Small & Medium Business
☐ MFG	
☒ TCD	☒ Research & Development

B.

Last Name Logan	Given First Name Beth	Nickname (if any) -	Middle Initial/Name T
Home Street Address 34 Springfield Street #2		Home Phone 617 629 3813	Pager Number -
City Somerville	State MA	Zip 02143	Citizenship Australia
Work Phone 617 551 7657	Work Fax 617 551 7650	Mail Code CRL	Employee # 331911
Name of Supervisor R. S. Nikhil		Name of Employer (if NOT an employee of Compaq) -	

GROUP	DIVISION
☐ CON	☐ Presario Mobile ☐ Peripherals ☐ Desktop ☐ Internet Services ☐ Con. Adv. Tech.
☐ ESSG	☐ Industry Standard Servers ☐ Storage Products ☐ Telecommunications ☐ High Perf. Server Bus. Unit ☐ Unix Bus. Unit ☐ NonStop™ e-Business ☐ Compaq Services ☐ Professional Services ☐ Alpha ☐ Tandem Bus. Unit
☐ CPCG	☐ Workstations ☐ Desktop ☐ Displays & Peripherals ☐ Portables ☐ Small & Medium Business
☐ MFG	

INVENTOR(S) SIGNATURE(S):

INVENTOR Date:	INVENTOR Date:
INVENTOR Date:	INVENTOR Date:

ADDITIONAL INVENTORS:

C.

Last Name Moreno	Given First Name Pedro	Nickname (if any)	Middle Initial/Name J
Home Street Address 6 Fort Washington Place		Home Phone 617 4992792	Pager Number
City Cambridge	State MA	Zip 02139	Citizenship SPAIN
Work Phone 617 551 7692	Work Fax 617 551 7650	Mail Code	Employee # 329353
Name of Supervisor R. S. Nikhil		Name of Employer (if NOT an employee of Compaq)	

GROUP	DIVISION
□ CON	□ Presario Mobile □ Peripherals □ Desktop □ Internet Services □ Con. Adv. Tech.
□ ESSG	□ Industry Standard Servers □ Storage Products □ Telecommunications □ High Perf. Server Bus. Unit □ Unix Bus. Unit □ NonStop™ e-Business □ Compaq Services □ Professional Services □ Alpha □ Tandem Bus. Unit
□ CPCG	□ Workstations □ Desktop □ Displays & Peripherals □ Portables □ Small & Medium Business
□ MFG	
☒ TCD	☒ Research & Development

D.

Last Name Suzek	Given First Name Baris	Nickname (if any)	Middle Initial/Name E
Home Street Address 3501 St Paul Street # 1106		Home Phone 410 261 5067	Pager Number
City Baltimore	State MD	Zip 21218	Citizenship Turkey
Work Phone 410 516 4650	Work Fax	Mail Code	Employee # 7744762
Name of Supervisor R. S. Nikhil		Name of Employer (if NOT an employee of Compaq) Student at Johns Hopkins University; all work relevant to this patent was performed while a summer intern at Compaq.	

GROUP	DIVISION
□ CON	□ Presario Mobile □ Peripherals □ Desktop □ Internet Services □ Con. Adv. Tech.
□ ESSG	□ Industry Standard Servers □ Storage Products □ Telecommunications □ High Perf. Server Bus. Unit □ Unix Bus. Unit □ NonStop™ e-Business □ Compaq Services □ Professional Services; □ Alpha □ Tandem Bus. Unit

□ CPCG	□ Workstations □ Desktop □ Displays & Peripherals □ Portables □ Small & Medium Business
□ MFG	
☒ TCD	☒ Research & Development

E.

Last Name	Given First Name	Nickname (if any)	Middle Initial/Name
Home Street Address		Home Phone	Pager Number
City	State	Zip	Citizenship
Work Phone	Work Fax	Mail Code	Employee #
Name of Supervisor		Name of Employer (if NOT an employee of Compaq)	

GROUP	DIVISION
□ CON	□ Presario Mobile □ Peripherals □ Desktop □ Internet Services □ Con. Adv. Tech.
□ ESSG	□ Industry Standard Servers □ Storage Products □ Telecommunications □ High Perf. Server Bus. Unit □ Unix Bus. Unit □ NonStop™ e-Business □ Compaq Services □ Professional Services □ Alpha □ Tandem Bus. Unit
□ CPCG	□ Workstations □ Desktop □ Displays & Peripherals □ Portables □ Small & Medium Business
□ MFG	
□ TCD	□ Research & Development

☒ TCD	☒ Research & Development

3. **CONCEPTION OF INVENTION:**

A. When did you first think of this invention? The invention grew from discussions which took place late June 2000.

B. Date of first written description? .June 2000

C. Please attach the first written description. (If submitting in electronic format, please scan all attachments and send).

D. If you can not send the first written description, please explain why and state where it can be found. -

E. Please list the name of others in Compaq to whom you've described the invention: A short presentation of the essential gist of this invention was given to interns and other Compaq Cambridge Research Laboratory employees during an "intern day" at Cambridge Research Laboratory on August 10 2000.

4. **IMPLEMENTING THE INVENTION**

A. Has the invention been implemented? ☒ Yes ☐ No ☐ Don't know
(Implementations can include physical prototypes, software, models, and simulations).

B. If implemented, please do not destroy, alter, or modify the implementation(s) without the authorization of the Compaq Legal Department, and answer the following questions for each implementation.

i. When was it implemented? June 15 2000 - July 31 (software)
ii. Where is the implementation now? (Attach or scan and send photograph, if possible)
iii. Has the implementation been tested? Yes
iv. If so, was the test successful? Yes

5. **USE OR SALE OF INVENTION:**

A. Has this been or will this be incorporated into a Compaq product? ☐ Yes ☒ No
If so, for each such product identify:

i. When was it or will it be incorporated into the product? ______
ii. Code name: ______
iii. Street name: ______

B. Has the invention been offered for sale or sold to anyone (e.g. an end user, vendor, reseller, partner, etc.)
☐ Yes ☒ No ☐ Don't know

i. If so, when: ______
ii. If so, to whom (name of company or individual): ______

C. If you don't know whether the invention has been offered for sale or sold, please provide the name of the best person to contact to determine when the invention has been or will be offered for sale or sold: Simon Kasif

NOTE: Please inform Compaq Legal immediately if, in the future, any of your answers under this Section 5 change so that we have ample opportunity to protect the invention within the time limits set out by law.

6. DISCLOSURE OF INVENTION TO OTHERS

A. Has a disclosure of the invention been made to any person(s) who is **NOT** a Compaq employee (including contractor, temporary, vendor, reseller, or partner and including conference presentations or journal articles)?
□ Yes ☒ No □ Don't know

B. If a disclosure was made, when was it made? ____________________

C. To whom was the disclosure made? ____________________

D. Was the disclosure made under an obligation of confidence? (e.g. Nondisclosure Agreement)
□ Yes □ No □ Don't know

7. DESCRIPTION OF THE INVENTION (continue on extra sheets as necessary)

A. To what type of technology does your invention relate? (Check all that apply)

CPU Technologies
- □ Keyboard/Mouse/Other Input Device
- □ Graphics
- □ Architecture
- □ Audio
- □ Memory
- □ Buses (ISA, EISA, PCI, AGP, other)
- □ Power Supplies/Batteries
- □ Other: ____________________

Peripherals Technologies
- □ Monitors/Screens
- □ CD-ROM
- □ DVD
- □ Tape Drives
- □ Disk Storage Systems
- □ Disk Controllers
- □ Printers
- □ Storage
- □ Other: ____________________

Other
- □ Manufacturing Processes
- □ Mechanical (functional)
- □ Mechanical (ornamental)
- □ PC/TV
- □ Racks
- ☒ Other: Algorithms for modeling proteins

Communications Technologies
- □ Network Interface Card
- □ Hubs/Concentrators
- □ Routers
- □ Switches
- □ Modems
- □ Remote Access
- □ Other: ____________________

Feature/Software Technologies
- □ Multiprocessor
- □ Fault Tolerance
- □ Remote Control
- □ Power Management
- □ Security
- □ Intelligent Manageability
- □ Smartstart
- □ Insight Manager
- □ Other ____________________

B. Describe the general subject matter of the invention. The invention addresses the problem of classifying, clustering or indexing proteins and other biological sequences such as genes by using an alternative representation based on high dimensional vectors. Each of the components of the vector represents the sensitivity of the protein (or sequence) to a particular biological motif (described later). Once obtained, this new representation can be used in conjunction with many exisiting machine learning techniques to analyse the sequences of interest. For example, this new representation can be combined with discriminative classification methods to classify new proteins from the amino acid sequence alone.

Computational methods for biological sequence analysis are playing an increasingly important role in biology and medicine. The key question addressed by these methods is the discovery of the function of a protein or gene. It is well known the function of a protein is dictated by its amino acid sequence since this determines the structure of the protein and thus its interaction with the environment.

Proteins are the building blocks of life, supporting a variety of functions which are essential for cell life. These include protection from infections or cancers, gene regulation, survival in different conditions, growth, differentiation, regeneration and others. In fact, the function of every cell in a living organism (whether microbial or human) is determined by which proteins (genes) are expressed in the cell and how they interact in the particular cell environment.

The area of protein function prediction is particularly timely because the new technology of high-throughput genomics generates thousands of hypothetical genes that have not been assigned a putative function. There are numerous commercial applications. Classifying new genes into categories opens many opportunities for new medical treatments. Genes are often used as drugs directly (e.g. insulin), or drug targets (e.g. attacking a particular gene in a microbial organisms). Other applications include the design of pesticides, design of new crops, gene therapies and rational drug design.

In this document we describe a new representation of proteins (genes) as objects in a very high-dimensional vector space. This representation offers numerous opportunities for predictive analysis of the space of biological sequences in a novel fashion deploying high-dimensional analysis techniques. The representation relies on aligning very short motif elements (biological templates) to the protein sequence. Subsequently, each protein is encoded as a multi-dimensional vector X, where dimension X_i corresponds to the score obtained by obtaining the maximum score of scoring (convolving) element E_i "against" the protein. The representation allows us to use existing templates (motifs) or "train" new ones.

C. Describe the particular problem faced by those working in the subject matter area. Proteins are macromolecules found in living organisms which play many roles essential to sustaining life (e.g. forming the physical framework of the organism, acting as enzymes to promote chemical reactions). A protein is composed of a sequence of several hundred amino acids. Proteins are created in living cells by translating the coding regions (genes) of the DNA sequence. Different proteins are expressed in different cells. The level of expression of different cells determines the cell function.

Since proteins are long and linear complex molecules, they `fold' to give a 3D shape. Biologists have identified 4 levels of structure which can influence the protein's function:
1. Primary Structure - the sequence of amino acids
2. Secondary Structure - the presence or absence of small `sub-folds'. These are regular patterns formed by local folding of the protein. (e.g. helices and sheets)
3. Tertiary Structure - the final 3D shape
4. Quaternary Structure - complexes formed with other proteins.
Given one level of structure, it is not necessarily a trivial task to predict the next level. Hence function prediction from the primary structure alone is difficult. Therefore, techniques other than sequencing are needed to determine the 3D structure and ultimately the protein function.

Lab-based experimental techniques to determine the tertiary structure are time-consuming and expensive or impossible for some proteins. This invention seeks to find a software-based solution.

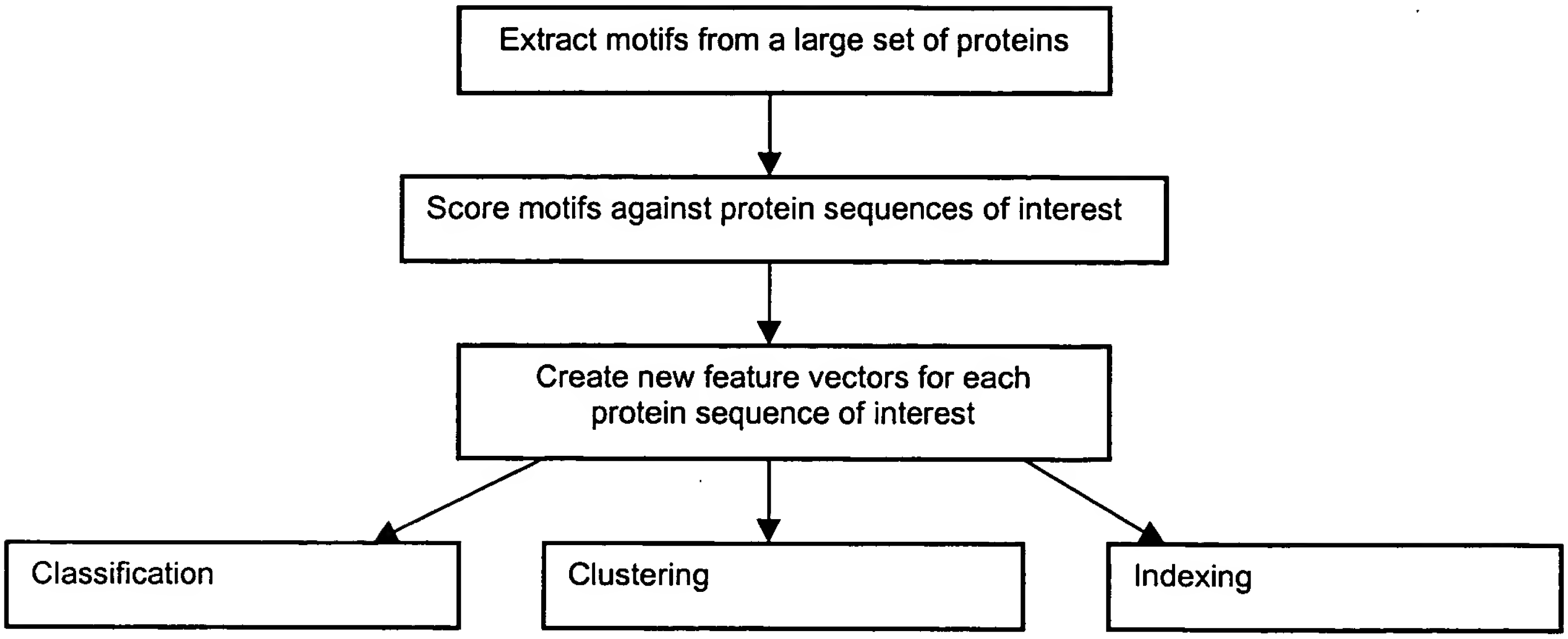
Currently, limited databases exist which contain protein domain sequences (primary structure) annotated with their secondary and tertiary structure. A protein domain is a subsequence of interest found in proteins. One use of this invention would be to use this labeled data to build models for known protein structures, and then to automatically annotate new proteins accordings to the models. However, the general idea of the invention could also to apply to other protein or gene classification problems and to cluster or index biological sequences.

D. Describe the old method(s) of performing the functions of the invention. The traditional and still most reliable way to perform protein structure prediction is to use laboratory-based techniques such as X-ray crystallography. However, recent years have seen the development of software-based solutions. One such technique is to use dynamic programming-based alignment tools such as `BLAST' to match the new sequence to previously labeled protein sequences (Altshul et all, 1990, Basic Local Alignment Search Tool, JMB 215:403-410). Alternatively, statistical techniques such as Hidden Markov Models (HMMs) can be use to build a model for each labeled class (E. Sonnhammer, S. Eddy and R. Durbin, Pfam: A Comprehensive Database of Protein Families Based on Seed Alignments, Proteins, 1997, 405—420.) (A. Krogh, M. Brown, I. Mian, K. Sjolander and D. Haussler, Hidden Markov Models in computational biology: Applications to protein modeling, J. of Molecular Biology, 1994, Volume 235, 1501--1531.). Still another alternative is to learn the boundaries between protein classes rather than a model for the class itself. (Jaakkola, Diekhans, Haussler (1999). Using the Fisher kernel method to detect remote protein homologies. Proceedings of ISMB'99). The first two approaches use the protein sequence itself directly to perform classification. The last one uses a HMM to compute the gradient of the protein being produced by the HMM with respect to each of the parameters of the HMM. In summary, none of these methods uses the sensitivity of parts of the protein to motifs to build a feature vector .

E. Why is the invention better than these old approaches? Lab-based techniques such as X-ray crystalographay are expensive and time-consuming. In addition, X-ray crystallography relies on having relatively large amounts of the protein. It cannot work with just a primary description of the protein (i.e. the sequence of amino acids in a file). Finally, it is not possible to crystallize certain proteins in any case (e.g, membrane spanning proteins).

BLAST and other dynamic programming methods are more time-consuming and less accurate than statistical-based techniques.

F. Attach at least one drawing or sketch of the invention if available.
(Attach or scan and send drawing or sketch)



Extract motifs from a large set of proteins
Score motifs against protein sequences of interest
Create new feature vectors for each protein sequence of interest
Classification
Clustering
Indexing

Figure 1. Overall algorithm of the invention.

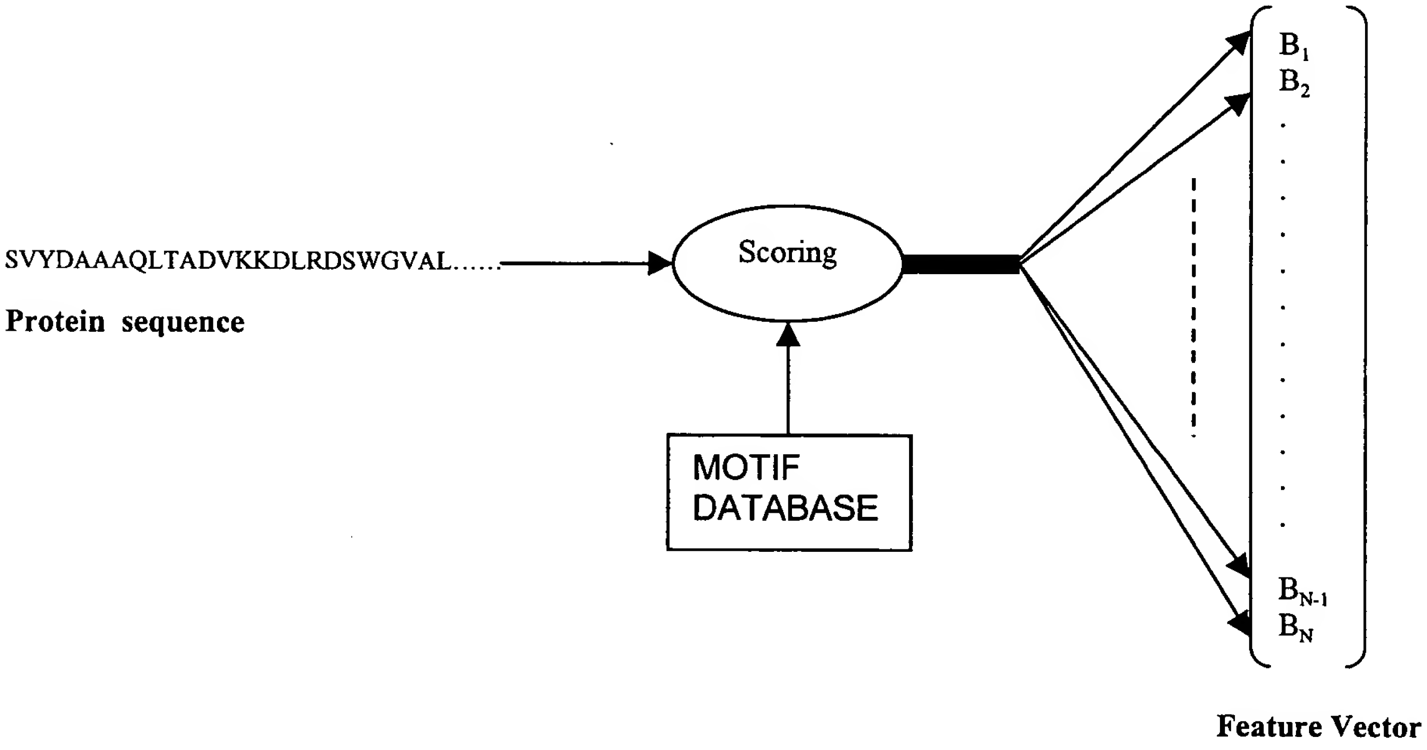


Figure 2: Process of creating feature vectors for each protein sequence of interest

G. Describe the invention, how it is used, and how it operates. Our process consists of 2 major steps. First we convert the amino acid sequences of interest to high dimensional feature vectors. Once this transformation has taken place, we can apply any number of statistical learning techniques to train models for classification, clustering or indexing the protein sequences. We describe these steps below. In this description, we shall describe the process as it applies to the analysis of protein sequences or subsequences. However, the technique could also be applied to DNA sequences or subsequences.

1. Creation of Feature Vectors

The first step converts each protein sequence or subsequence of interest to a new representation of fixed length, i.e. any protein sequence no matter how long it is, is converted into a feature vector of fixed length. Each dimension of these feature vectors represents the sensitivity of the protein to a particular biological motif. Therefore, in order to create feature vectors, we first create or obtain a database of short, highly conserved regions in related protein domains. Such regions are often called called `blocks', `motifs' or `probabilistic templates'.

A motif can be represented by a K by L matrix M in which each of the K rows represents a particular amino acid (or nucleotide for DNA sequences) and L represents the length of the motif. For protein sequences, K = 20. (For DNA sequences K = 4.) Each cell M[amino acid,position] in the matrix represents the probability of that amino acid in that position. This matrix can also store log-ratios rather than probabilities. Thus a motif can be thought of as a 0-th order Markov model. A motif of length L is scored against a protein by computing the probability of every subsequence of length L in the protein being generated by the model that corresponds to the motif.

The BLOCKS database (Steven Henikoff & Jorja G. Henikoff, "Automated assembly of protein blocks for database searching", Nucleic Acids Research, 19:23, p. 6565-6572. 1991) is an example of a database of motifs. Emotif (http://dna.stanford.edu/emotif/), and PRINTs (http://bioinf.man.ac.uk/dbbrowser/PRINTS/) are other such databases. These could all be used in our invention. Alternatively, it is possible to create a new motif database from any protein database which has been labeled according to some parameter (e.g. structure). This is achieved by using multiple alignment software to find short mulitply aligned ungapped sequences and then collecting statistics about these in a matrix (http://www2.ebi.ac.uk/clustalw/, http://www.blocks.fhcrc.org/). By creating a motif database specific to the proteins of interest, more meaningful feature vectors may be obtained since the motifs from a more general database may not occur in the proteins of interest.

To create a feature vector for each protein sequence of interest we search for each motif in the sequence as described above. The result is an N-dimensional feature vector where N is the total number of motifs in our database. Each dimension J contains a score describing the degree of aligment of motif J to the protein domain. For the case where a motif is detected multiple times in a domain, we can apply a variety of heuristics. For example, we can take the maximum of all scores for that block in that domain or the sum of such scores. In preliminary experiments, we found that taking the maximum score gives superior classification peformance. We can also apply a threshold such that scores below a certain number are set to zero. Additionally, given the complete set of feature vectors for all protein domains in the training set, we can reduce the dimensionality of these vectors using standard dimension reduction techniques such as Principal Components Analysis (PCA).

2. Clustering, Classification and Indexing

Once all the protein sequences or subsequences of interest have been transformed to feature vectors, models can be generated to describe these features and perform clustering, classification or indexing. We describe each of these processes below.

2.a. Clustering

Clustering groups together proteins with similar feature vectors in order to discover previously unknown relationships between them. For example, using well known algorithms such as k-means or nearest neighbors it is possible to decide if two proteins as represented by our new feature vector are close or not. The key concept here is that the new representation allows us to compare proteins both reliably and effectively.

2.b. Classification

Classification attempts to learn a relationship or model given a set of labeled feature vectors called the `training set'. Each label denotes the class that the vector belongs to. For example, the classes may defined by protein structural information. Possibly the labeling is generated by clustering. Given this model, unseen vectors, usually denoted the `testing set', are assigned labels according to the models learnt. An example of the classification of proteins into structural classes is described below.

2.c. Indexing

Indexing organizes a database of protein sequences in such a way that for a given protein (represented by its feature vector), `similar' proteins can be found efficiently. A possible implementation would be to use the NI2 index to index a database of proteins as represented with our new proposed high dimensional representations. A new "query" protein can be presented to NI2 and all similar proteins can be retrieved. The similarity function used in NI2 would need to be changed and many possibilies exist. Clustering and classification techniques usually form an integral part of indexing algorithms. The main idea here is to use the index to retrieve the most similar proteins to a given query. This operation has important applications for biologists that are involved in drug design since a set of similar proteins can suggest multiple possible functions for a given query proteins. Rather than a single classification into a single structural class.

H. Describe the construction and structure of the preferred implementation of the invention. We have implemented a system which can classify protein domains according to their tertiary structure.

Our process consists of 4 steps.

1. Given a set of training protein domains labeled according to structure, convert each of these into a multidimensional feature vector as described above. We use hits from the BLOCKs motif database to create our vectors.

2. Given the labeled feature vectors, we learn Support Vector Machine (SVM)

classifiers (Burger, 1998, "A tutorial on Support Vector Machines for Pattern Recognition", Data Mining and Knowledge Discovery Journal) to separate each structural class from "the rest of the world". A SVM classifier learns a separating hyperplane between two classes which maximises the `margin' - the distance between the hyperplane and the nearest datapoint of each class. The appeal of SVMs is twofold. First they do not require any complex tuning of parameters, and second they exhibit a great ability to generalize give a small training corpora. They are particularly amenable for learning in high dimensional spaces. The only parameters needed to tune a SVM are the `capacity' and the choice of kernel. The capacity allows us to control how much tolerance for errors in the classification of training samples we allow and therefore the generalization ability of the SVM. A SVM with high capacity will classify all training samples correctly but will not be able to generalize well for testing samples. In effect, it will construct a classifier too tuned for the training samples which will limit its ability to generalize later on when testing samples are presented to the system. Conversely, a very low capacity will produce a classifier that does not fit the data suffiiciently accurately. It will allow many training and testing samples to be classified incorrectly.

The second tuning parameter is the kernel. The kernel function allows the SVM to create hyperplanes in high dimensional spaces that effectively separate the training data. Often in the input space training vectors cannot be separated by a simple hyperplane. The kernel allows transforming the data from one space to another space where a simple hyperplane can effectively separate the data in two classes.

We tune these two parameters separately for each structural family of interest.

An additional step consists of tuning the operating point of the classifier so that we can control the amount of false negatives. In our implementation we find a threshold value such that any score returned by the SVM that is bigger than this guarantees no false negatives.

3. Given a set of unlabeled structural domains (the testing set) we convert each of these vectors to a multidimensional feature vector using BLOCKS as before.

4. Now, for each unlabeled feature vector, to determine if it belongs to a particular class we test it using the SVM created for that class. The SVM classifier will produce a "score" representing the distance of the testing feature vector from the margin. The bigger the score the further away the vector is from the margin and the more confident the classifier is in its own output. If the score is below the threshold set in Step 2, we classify the vector (and hence the corresponding protein) as belonging to that particular class. Otherwise, it is classified as not belonging to the class.
For multi-class classification we can use standard procedures such as classifying based on the highest score returned by each of the individual classifiers.

I. Is the invention designed to conform or enhance any industry standard?
□ Yes ☒ No □ Don't Know
If so, what industry standard? ________________________________

COPY

Peg Norcutt

From:	Lange, Rich [rich.lange@compaq.com]
Sent:	Friday, September 15, 2000 9:48 PM
To:	Mary Lou Wakimura (E-mail)
Cc:	Kasif, Simon; Logan, Beth; Strong, Diane; Munson, Susan
Subject:	RE: CR Filing Approval- P00-3373 - Technique for Protein and Gene Cla ssification /Clustering/Indexing via a Fixed Dimensional Vector..

By 12/15/00

MaryLou,
Per you voice mail message, your quote of $12-14k for preparation and filing this case in the PTO is approved. Please contact the inventors and proceed.
Thanks.
Rich

-----Original Message-----
From: Lange, Rich
Sent: Friday, September 15, 2000 11:38 AM
To: 'marylou.wakimura@hbsr.com'
Subject: FW: CR Filing Approval- P00-3373 - Technique for Protein and Gene Cla ssification /Clustering/Indexing via a Fixed Dimensional Vector..

-----Original Message-----
From: Lange, Rich
Sent: Friday, September 15, 2000 11:36 AM
To: 'Wakimura, MaryLou'; 'Smith, Jim'
Subject: FW: CR Filing Approval- P00-3373 - Technique for Protein and Gene Cla ssification /Clustering/Indexing via a Fixed Dimensional Vector..

MaryLou/Jim
Please call me to discuss the capability of your firm to prepare and file the attached case.
Thanks.
Rich

-----Original Message-----
From: Logan, Beth
Sent: Thursday, September 07, 2000 8:38 AM
To: Lange, Rich
Subject: FW: CR Filing Approval- P00-3373 - Technique for Protein and Gene Cla ssification /Clustering/Indexing via a Fixed Dimensional Vector..

Rich
Here is the `Protein' patent disclosure.
Beth

-----Original Message-----
From: Reed, Bob [mailto:Bob.Reed@compaq.com]
Sent: Tuesday, September 05, 2000 8:26 AM
To: Kasif, Simon; Logan, Beth; Moreno, Pedro
Cc: Nikhil, Rishiyur S; Jouppi, Norm; Lange, Rich; Ulichney, Bob; Williams, Eric; Burrows, Mike; Iannucci, Bob; Munson, Susan; Strong,

EXHIBIT
I

Diane
Subject: CR Filing Approval- P00-3373 - Technique for Protein and Gene Cla ssification /Clustering/Indexing via a Fixed Dimensional Vector..

Dear Inventors,

Your approved IDF has been submitted to our CPQ/CR patent law team for counsel assignment. Rich Lange and Sue Munson of CPQ Law West will be supporting you during the application process.

Regards,
Bob Reed
CR PRC

Docket# P00-3373

Status: APR - Approved - Not Commissioned

-----Original Message-----
From: Reed, Bob
Sent: Monday, August 21, 2000 3:19 PM
To: Jouppi, Norm; Lange, Rich; Ulichney, Bob; Williams, Eric (LKG); Burrows, Mike
Cc: Kasif, Simon; Logan, Beth; Moreno, Pedro; Nikhil, Rishiyur S; Iannucci, Bob
Subject: Invention Review Rq - Technique for Protein and Gene Classification /Clustering/Indexing via a Fixed Dimensional Vector..

Dear CR Invention Review Committee Members,

Please review the attached invention disclosure and reply with your comments and recommendations for filing to: bob.reed@compaq.com

Our target submission date of an approved IDF to CPQ Law is: September 5, 2000.

TITLE: Technique for Protein and Gene Classification /Clustering/Indexing via a Fixed Dimensional Vector...

INVENTORS: Kasif, Logan, Moreno, Suzek

LAB: CRL

STATUS: IDR - Invention Disclosure Received

Thank you for your prompt attention regarding this matter.

Regards,

Bob Reed

CR IR Committee
Mike Burrows, SRC
Norm Jouppi, WRL
Bob Ulichney, CRL
Eric Williams, CSG
Rich Lange, Law

Bob Reed, CR
cc:
Bob Iannucci, CR

-----Original Message-----
From: Logan, Beth
Sent: Monday, August 21, 2000 2:14 PM
To: Reed, Bob
Cc: Logan, Beth; Moreno, Pedro; Simon Kasif
Subject: Patent disclosure - Technique for Protein and Gene
Classification /Clustering/Indexing via a Fixed Dimensional Vector
formed using Alignmen ts of Small Motif or Blocks

Bob,

Please find enclose the patent disclosure for "Technique for Protein and Gene Classification/Clustering/Indexing via a Fixed Dimensional Vector formed using Alignments of Small Motif or Blocks" (Kasif/Logan/Moreno/Suzek). As the first written description of the subject matter, I attach the web page from the relevant student project. Since I believe this is one of the first if not the first computational biology patent to be sent to the committee, please do not hesitate to contact us if the subject matter is so unfamiliar that it is difficult to understand.

Beth Logan   Email: Beth.Logan@compaq.com

Compaq Computer Corporation Ph: +1 617 551 7657
One Cambridge Center   Fax: +1 617 551 7650
Cambridge MA 02142 USA   WWW: http://www.crl.research.digital.com
<http://www.crl.research.digital.com>

COPY

Peg Norcutt

From: Beth Logan [btl@crl.dec.com]
Sent: Friday, September 22, 2000 10:45 AM
To: 'MaryLou.Wakimura@hbsr.com'
Subject: RE: CR Filing Approval- P00-3373 - Technique for Protein and Gene Cla ssification /Clustering/Indexing via a Fixed Dimensional Vector..

Hi Mary Lou
How about Thursday October 5 at 11am here at CRL ?
Beth

> -----Original Message-----
> From: MaryLou.Wakimura@hbsr.com [mailto:MaryLou.Wakimura@hbsr.com]
> Sent: Thursday, September 21, 2000 9:00 PM
> To: btl@crl.dec.com
> Subject: RE: CR Filing Approval- P00-3373 - Technique for Protein and
> Gene Cla ssification /Clustering/Indexing via a Fixed Dimensional
> Vector..
>
>
>
> Hi Beth
> Would you be available Tues Oct 3 around 10a or Thur OCt 5
> (up until 2p)?
> Either of these would work for me to come to your office.
> Just let me know
> Thanks
> --Mary Lou 781- 861-6240 x3214
> -----Original Message-----
> From: Beth Logan
> To: 'marylou.wakimura@hbsr.com'
> Cc: Beth Logan
> Sent: 9/21/00 5:11 PM
> Subject: RE: CR Filing Approval- P00-3373 - Technique for
> Protein and Gene
> Cla ssification /Clustering/Indexing via a Fixed Dimensional Vector..
>
> MaryLou
> Could you please give an estimate as to when we can meet
> regarding this
> patent. I will be away from Boston at conferences from 12
> October - 25
> October inclusive and it would be good if we could meet before then to
> get
> things started.
> Yours
> Beth
> --
> Beth Logan Email: Beth.Logan@compaq.com
>
> Compaq Computer Corporation Ph: +1 617 551 7657
> One Cambridge Center Fax: +1 617 551 7650
> Cambridge MA 02142 USA WWW:
> http://www.crl.research.digital.com
>
>
> > -----Original Message-----
> > From: Lange, Rich [mailto:rich.lange@compaq.com]
> > Sent: Friday, September 15, 2000 9:44 PM

EXHIBIT
J

> > To: Lange, Rich
> > Cc: Kasif, Simon; Logan, Beth; Strong, Diane; Munson, Susan
> > Subject: RE: CR Filing Approval- P00-3373 - Technique for
> Protein and
> > Gene Cla ssification /Clustering/Indexing via a Fixed Dimensional
> > Vector..
> >
> >
> > MaryLou,
> > Per you voice mail message, your quote of $12-14k for
> > preparation and filing
> > this case in the PTO is approved. Please contact the
> > inventors and proceed.
> > Thanks.
> > Rich
> >
> > -----Original Message-----
> > From: Lange, Rich
> > Sent: Friday, September 15, 2000 11:38 AM
> > To: 'marylou.wakimura@hbsr.com'
> > Subject: FW: CR Filing Approval- P00-3373 - Technique for
> Protein and
> > Gene Cla ssification /Clustering/Indexing via a Fixed Dimensional
> > Vector..
> >
> >
> >
> >
> > -----Original Message-----
> > From: Lange, Rich
> > Sent: Friday, September 15, 2000 11:36 AM
> > To: 'Wakimura, MaryLou'; 'Smith, Jim'
> > Subject: FW: CR Filing Approval- P00-3373 - Technique for
> Protein and
> > Gene Cla ssification /Clustering/Indexing via a Fixed Dimensional
> > Vector..
> >
> >
> > MaryLou/Jim
> > Please call me to discuss the capability of your firm to
> > prepare and file
> > the attached case.
> > Thanks.
> > Rich
> >
> > -----Original Message-----
> > From: Logan, Beth
> > Sent: Thursday, September 07, 2000 8:38 AM
> > To: Lange, Rich
> > Subject: FW: CR Filing Approval- P00-3373 - Technique for
> Protein and
> > Gene Cla ssification /Clustering/Indexing via a Fixed Dimensional
> > Vector..
> >
> >
> > Rich
> > Here is the `Protein' patent disclosure.
> > Beth
> >
> > -----Original Message-----
> > From: Reed, Bob [mailto:Bob.Reed@compaq.com]
> > Sent: Tuesday, September 05, 2000 8:26 AM
> > To: Kasif, Simon; Logan, Beth; Moreno, Pedro
> > Cc: Nikhil, Rishiyur S; Jouppi, Norm; Lange, Rich; Ulichney, Bob;

> > Williams, Eric; Burrows, Mike; Iannucci, Bob; Munson, Susan; Strong,
> > Diane
> > Subject: CR Filing Approval- P00-3373 - Technique for
> Protein and Gene
> > Cla ssification /Clustering/Indexing via a Fixed
> Dimensional Vector..
> >
> >
> > Dear Inventors,
> >
> > Your approved IDF has been submitted to our CPQ/CR patent
> law team for
> > counsel assignment. Rich Lange and Sue Munson of CPQ Law
> West will be
> > supporting you during the application process.
> >
> > Regards,
> > Bob Reed
> > CR PRC
> >
> > Docket# P00-3373
> >
> > Status: APR - Approved - Not Commissioned
> >
> >
> >
> >
> >
> >
> > -----Original Message-----
> > From: Reed, Bob
> > Sent: Monday, August 21, 2000 3:19 PM
> > To: Jouppi, Norm; Lange, Rich; Ulichney, Bob; Williams, Eric
> > (LKG); Burrows,
> > Mike
> > Cc: Kasif, Simon; Logan, Beth; Moreno, Pedro; Nikhil,
> > Rishiyur S; Iannucci,
> > Bob
> > Subject: Invention Review Rq - Technique for Protein and Gene
> > Classification
> > /Clustering/Indexing via a Fixed Dimensional Vector..
> >
> >
> > Dear CR Invention Review Committee Members,
> >
> > Please review the attached invention disclosure and reply
> > with your comments
> > and recommendations for filing to: bob.reed@compaq.com
> >
> > Our target submission date of an approved IDF to CPQ Law is:
> > September 5,
> > 2000.
> >
> > TITLE: Technique for Protein and Gene Classification
> > /Clustering/Indexing
> > via a Fixed Dimensional Vector...
> >
> > INVENTORS: Kasif, Logan, Moreno, Suzek
> >
> > LAB: CRL
> >
> > STATUS: IDR - Invention Disclosure Received
> >
> > Thank you for your prompt attention regarding this matter.

> >
> > Regards,
> >
> > Bob Reed
> >
> > CR IR Committee
> > Mike Burrows, SRC
> > Norm Jouppi, WRL
> > Bob Ulichney, CRL
> > Eric Williams, CSG
> > Rich Lange, Law
> > Bob Reed, CR
> > cc:
> > Bob Iannucci, CR
> >
> >
> > -----Original Message-----
> > From: Logan, Beth
> > Sent: Monday, August 21, 2000 2:14 PM
> > To: Reed, Bob
> > Cc: Logan, Beth; Moreno, Pedro; Simon Kasif
> > Subject: Patent disclosure - Technique for Protein and Gene
> > Classification /Clustering/Indexing via a Fixed Dimensional Vector
> > formed using Alignmen ts of Small Motif or Blocks
> >
> >
> > Bob,
> >
> > Please find enclose the patent disclosure for "Technique for
> > Protein and
> > Gene Classification/Clustering/Indexing via a Fixed
> Dimensional Vector
> > formed using Alignments of Small Motif or Blocks"
> > (Kasif/Logan/Moreno/Suzek). As the first written description
> > of the subject
> > matter, I attach the web page from the relevant student
> > project. Since I
> > believe this is one of the first if not the first
> > computational biology
> > patent to be sent to the committee, please do not hesitate to
> > contact us if
> > the subject matter is so unfamiliar that it is difficult to
> > understand.
> >
> > Beth Logan Email: Beth.Logan@compaq.com
> >
> > Compaq Computer Corporation Ph: +1 617 551 7657
> > One Cambridge Center Fax: +1 617 551 7650
> > Cambridge MA 02142 USA WWW:
> http://www.crl.research.digital.com
> <http://www.crl.research.digital.com>
>